Language of Models

Dr. Maria Tackett

10.10.19

Click for PDF of slides

Announcements

- Labs resume on Friday
- Solutions to webscraping application exercise

The language of models

Modeling

- Use models to explain the relationship between variables and to make predictions
- For now we focus on **linear** models (but remember there are other types of models too!)

Packages

Packages

- You're familiar with the tidyverse:

```
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.

```
library(broom)
```

Data: Paris Paintings

Paris Paintings

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

What does the **`data/`** mean in the code above? Hint: Where is the data file located?

- Paris Paintings Codebook

Meet the data curators

Sandra van Ginhoven Hilary Coe Cronheim

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

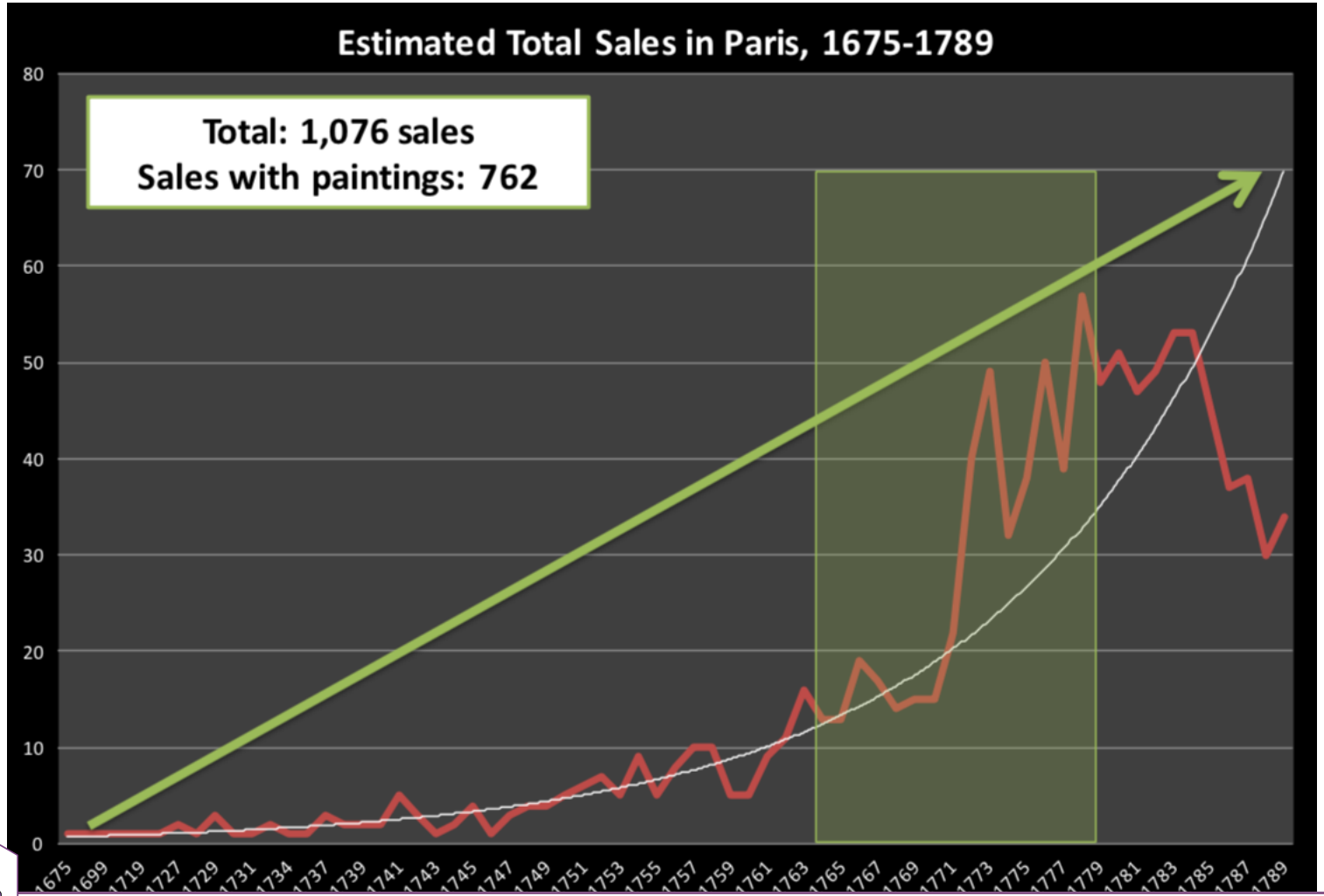
- Source: Printed catalogues of 28 auction sales in Paris, 1764- 1780
- 3,393 paintings, their prices, and descriptive details from sales catalogues over 60 variables

Auctions today

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

Auctions back in the day

Paris auction market

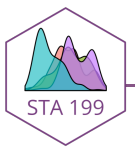


Depart pour la chasse

Auction catalogue text

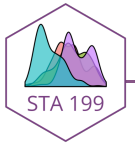
```
pp %>% filter(name == "R1777-89a") %>%
  select(name:endbuyer) %>% t()
```

```
##                   [,1]
## name              "R1777-89a"
## sale              "R1777"
## lot               "89"
## position          "0.3755274"
## dealer            "R"
## year              "1777"
## origin_author     "D/FL"
## origin_cat        "D/FL"
## school_pntg       "D/FL"
## diff_origin       "0"
## logprice          "8.575462"
## price             "5300"
## count             "1"
## subject           "D\x8epart pour la chasse"
## authorstandard    "Wouwerman, Philips"
## artistliving      "0"
## authorstyle       NA
## author            "Philippe Wouwermans"
## winningbidder     "Langlier, Jacques for Poullain, Antoine"
## winningbiddertype "DC"
## endbuyer          "C"
```



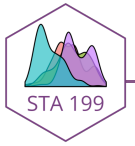
```
pp %>% filter(name == "R1777-89a") %>%
  select(Interm:finished) %>% t()
```

```
##                [,1]
## Interm         "1"
## type_intermed "D"
## Height_in      "17.25"
## Width_in       "23"
## Surface_Rect   "396.75"
## Diam_in        NA
## Surface_Rnd    NA
## Shape          "squ_rect"
## Surface        "396.75"
## material       "bois"
## mat            "b"
## materialCat    "wood"
## quantity       "1"
## nfigures       "0"
## engraved       "0"
## original       "0"
## prevcoll       "1"
## othartist      "0"
## paired         "1"
## figures        "0"
## finished       "0"
```



```
pp %>% filter(name == "R1777-89a") %>%
  select(lrgfont:other) %>% t()
```

```
##             [,1]
## lrgfont        0
## relig          0
## landsALL       1
## lands_sc       0
## lands_elem     1
## lands_figs     1
## lands_ment     0
## arch           1
## mytho          0
## peasant        0
## othgenre       0
## singlefig      0
## portrait       0
## still_life     0
## discauth       0
## history        0
## allegory       0
## pastorale      0
## other          0
```

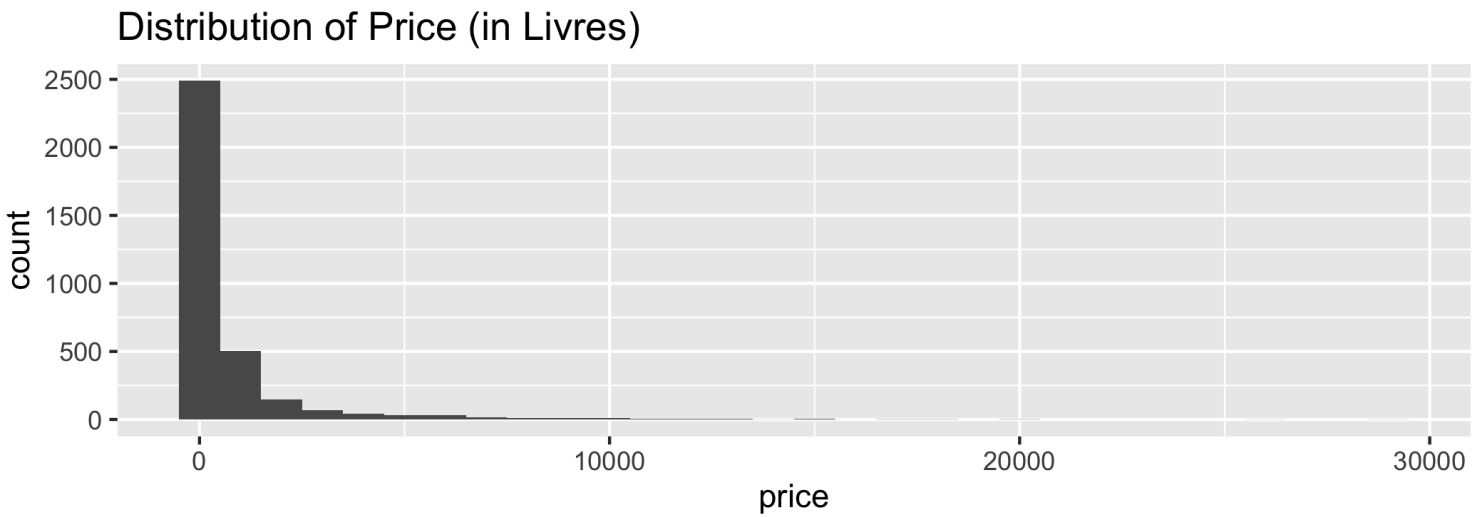


Modeling the relationship between variables

Prices

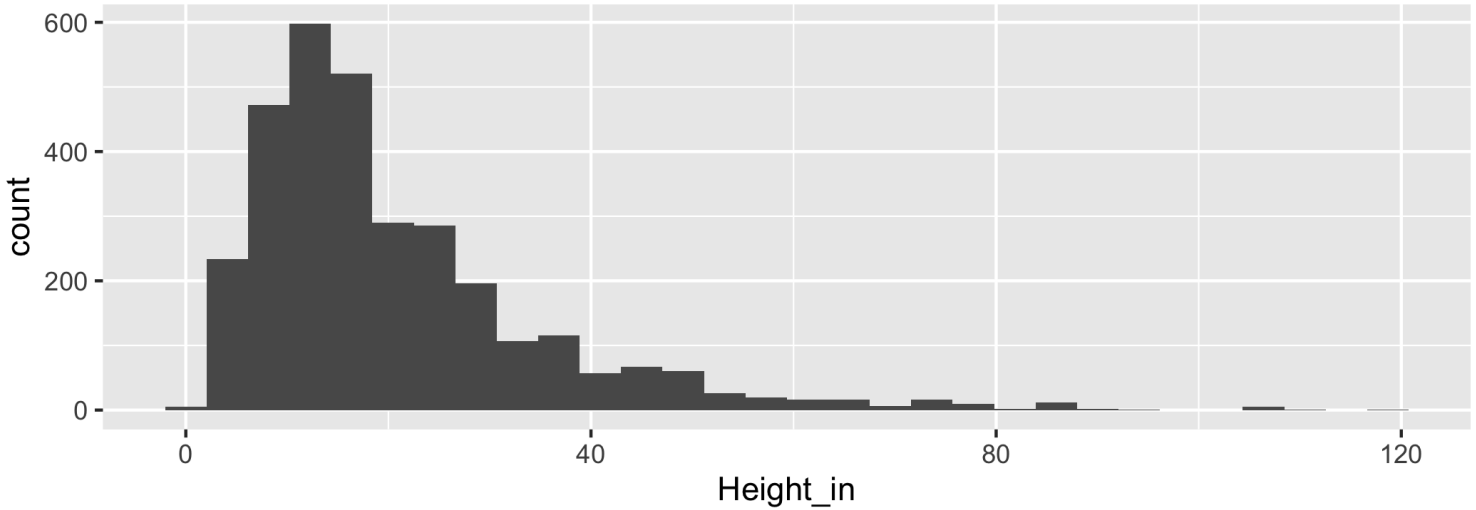
Describe the distribution of prices of paintings.

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



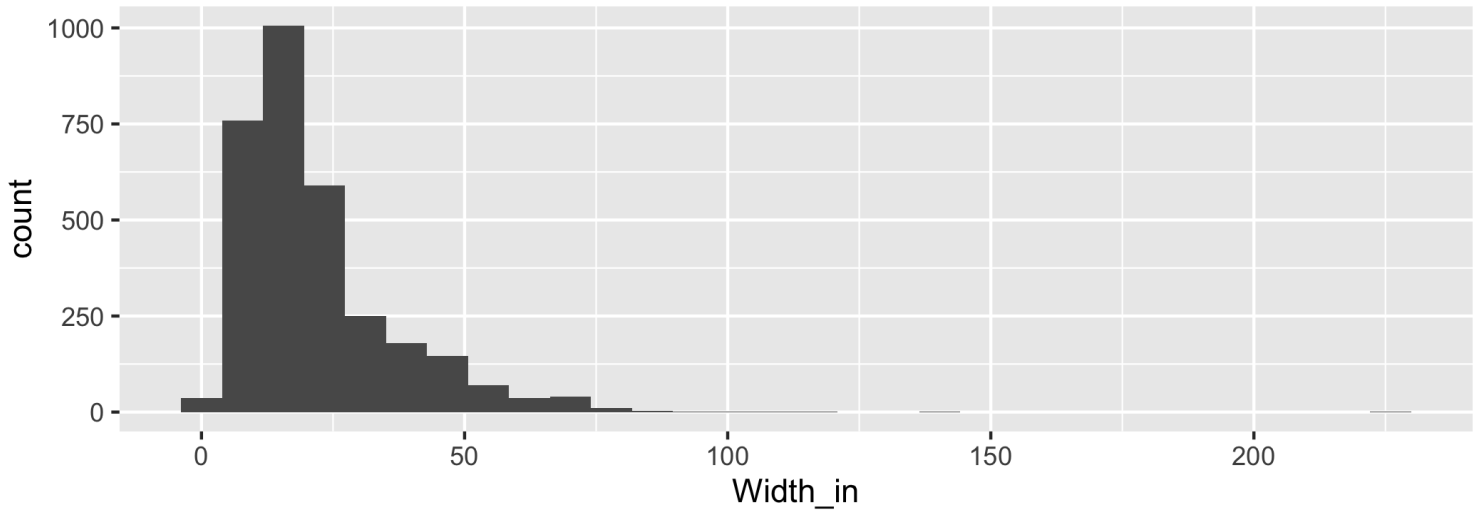
Height

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



Width

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



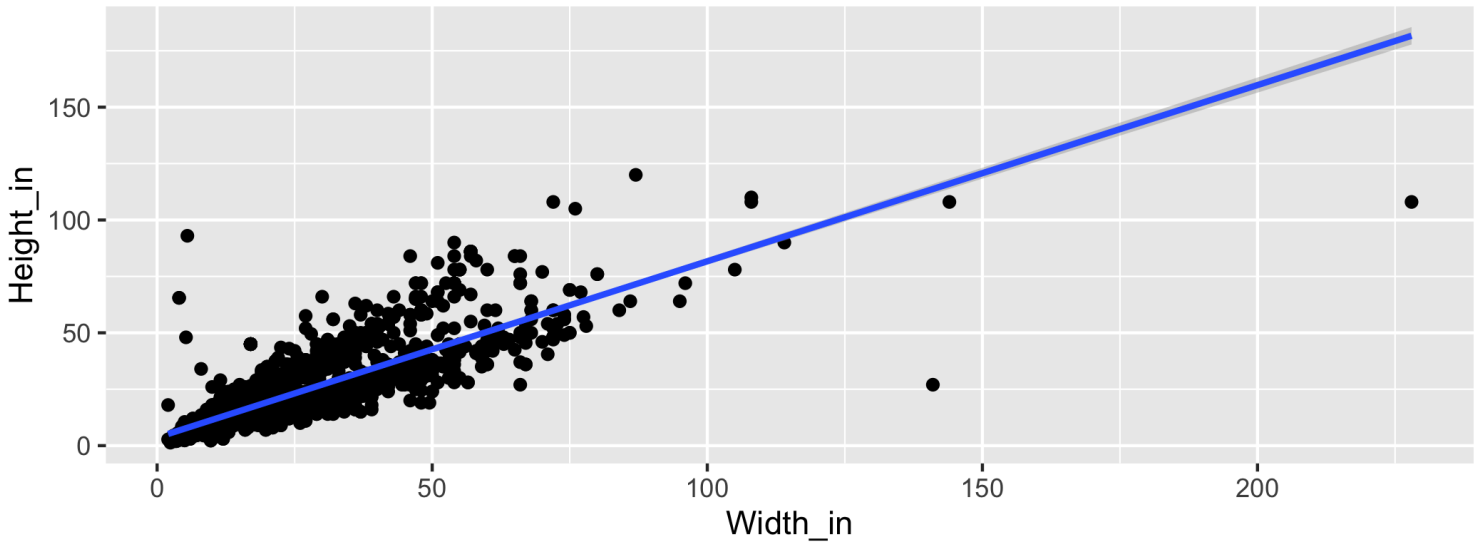
Models as functions

- We can represent relationships between variables using **functions**
- A function is a mathematical concept: the relationship between an output and one or more inputs.
  - Plug in the inputs and receive back the output
- Example: 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
```

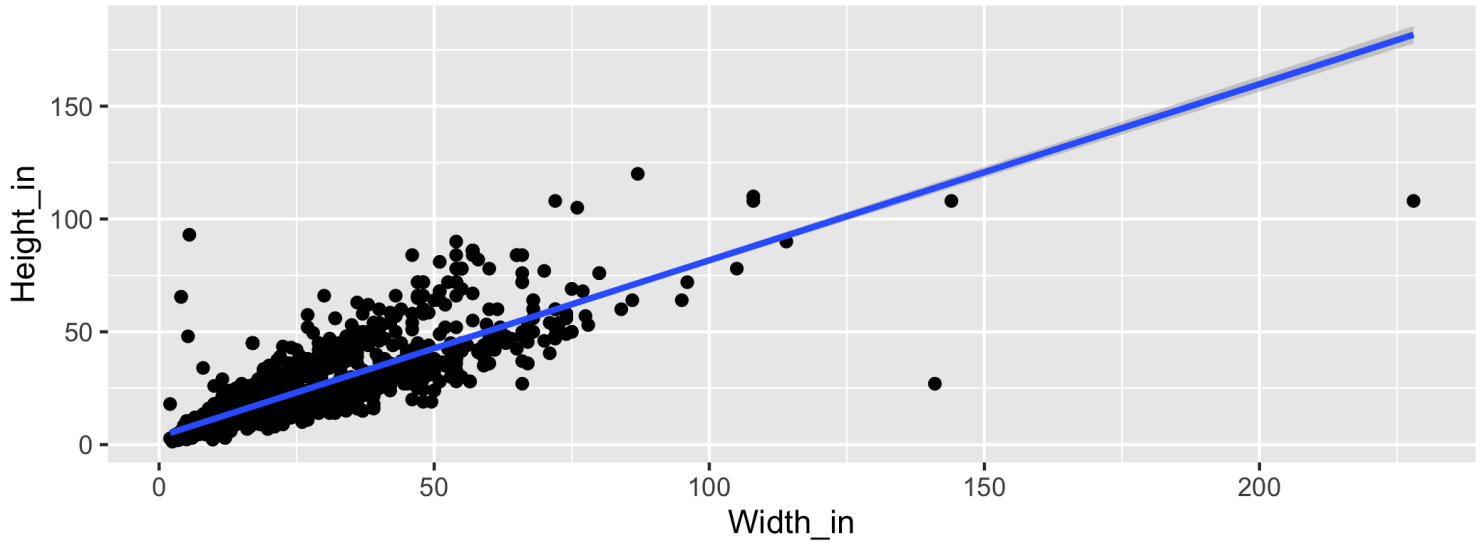
Height as a function of width

Describe the relationship between height and width of paintings.



Visualizing the linear model

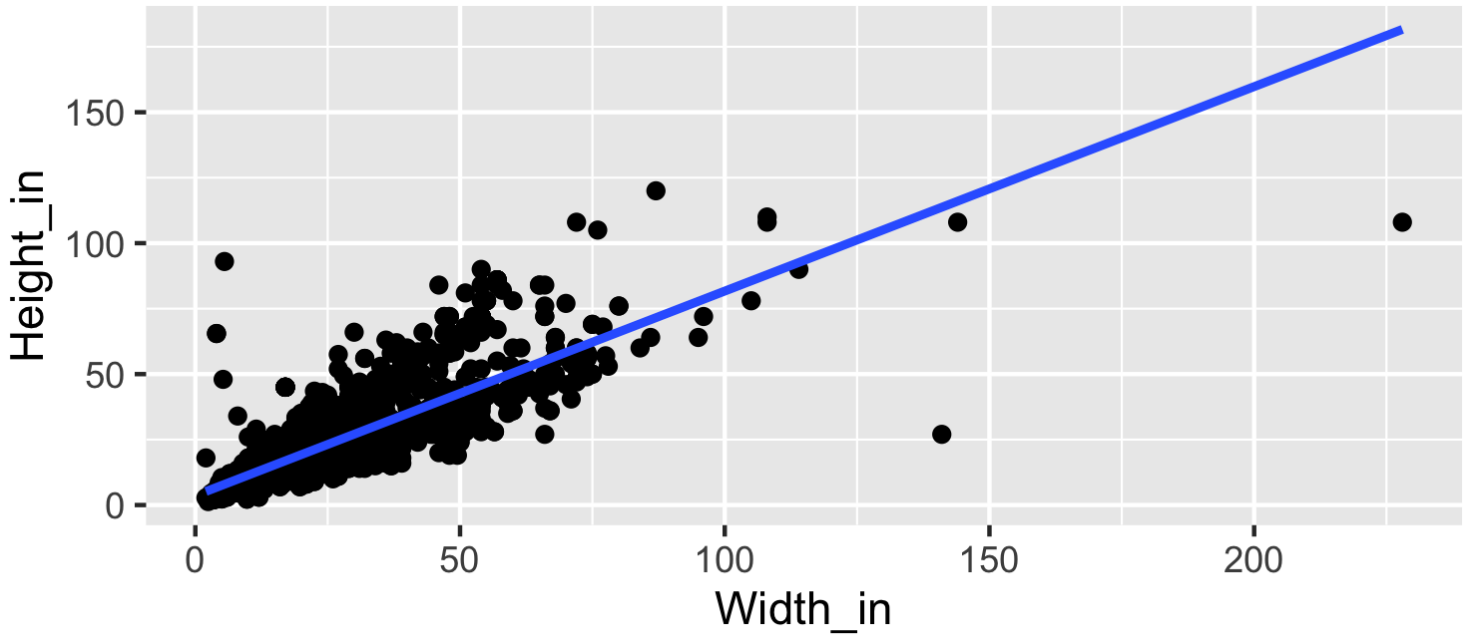
```
ggplot(data = pp, aes(x = Width_in, y = Height_in)) +
  geom_point() +
  geom_smooth(method = "lm") # lm for linear model
```



Visualizing the linear model

... without the measure of uncertainty around the line

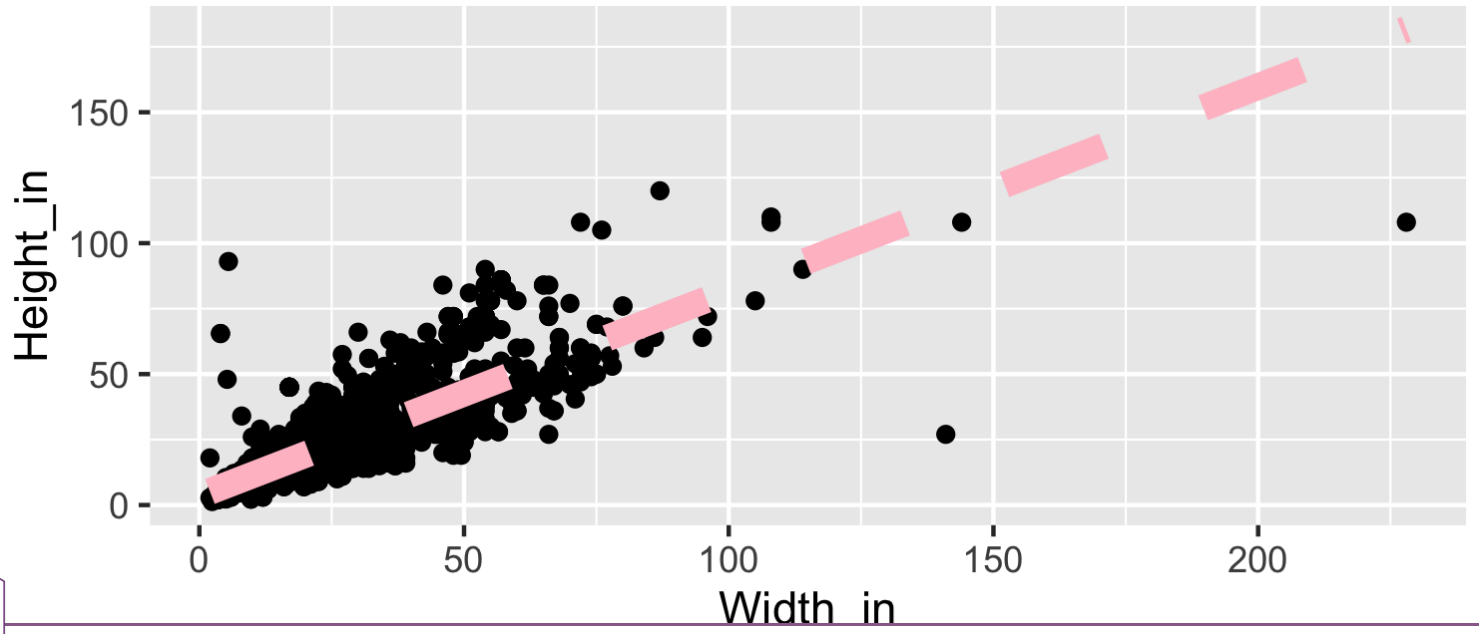
```
ggplot(data = pp, aes(x = Width_in, y = Height_in)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) # lm for linear model
```



Visualizing the linear model

... with different cosmetic choices for the line

```
ggplot(data = pp, aes(x = Width_in, y = Height_in)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE,
              # color       #line type  #line weight
              col = "pink", lty = 2,    lwd = 3)
```

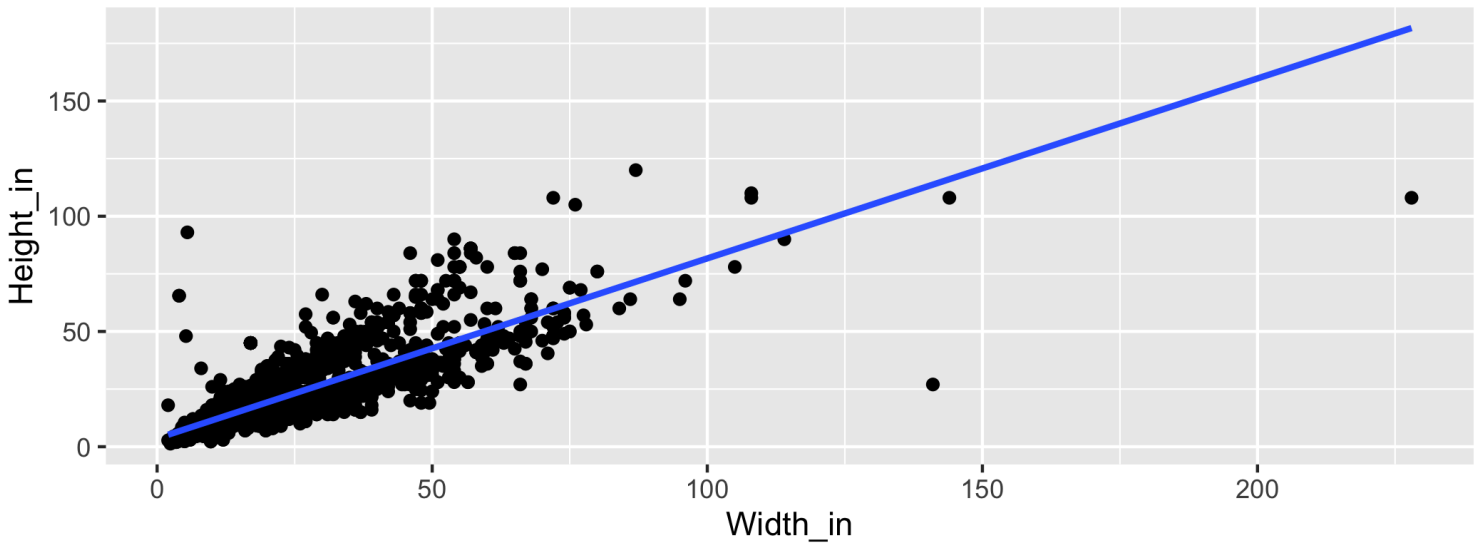


Vocabulary

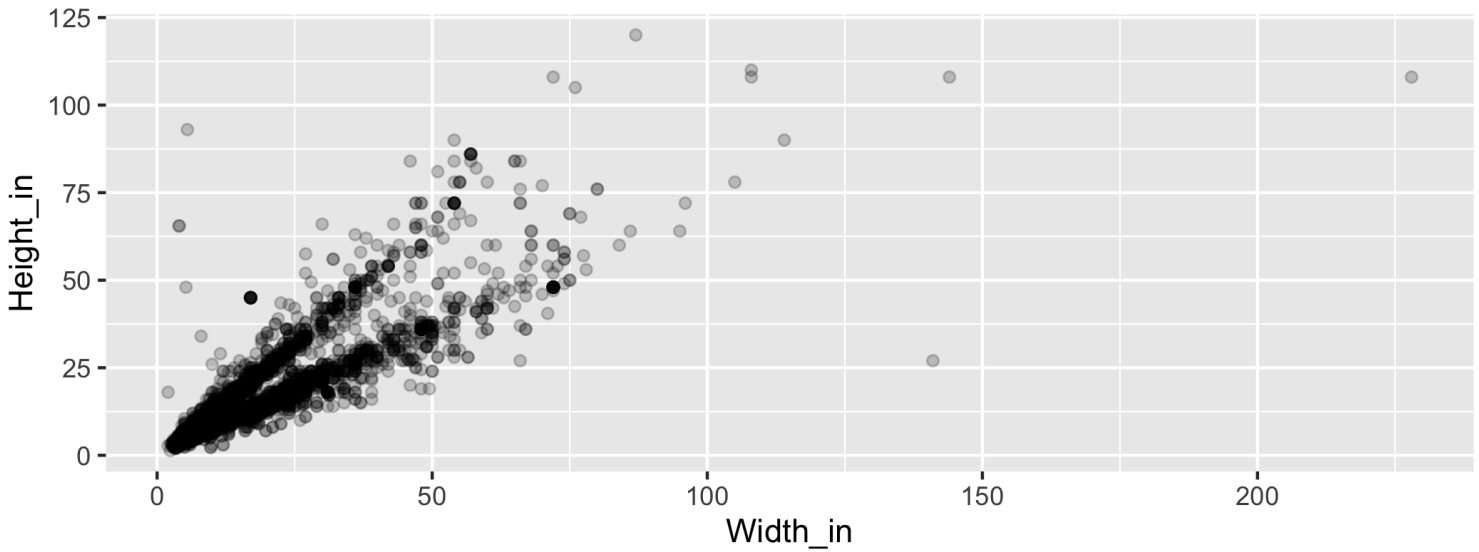
- **Response variable:** Variable whose behavior or variation you are trying to understand, on the y-axis (dependent variable)
- **Explanatory variables:** Other variables that you want to use to explain the variation in the response, on the x-axis (independent variables)
- **Predicted value:** Output of the **model function**
  - The model function gives the typical value of the response variable *conditioning* on the explanatory variables
- **Residuals:** Show how far each case is from its model value
  - **Residual = Observed value - Predicted value**
  - Tells how far above/below the model function each case is

Residuals

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



The plot below displays the relationship between height and width of paintings. It uses a lower alpha level for the points than the previous plots we looked at. 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.

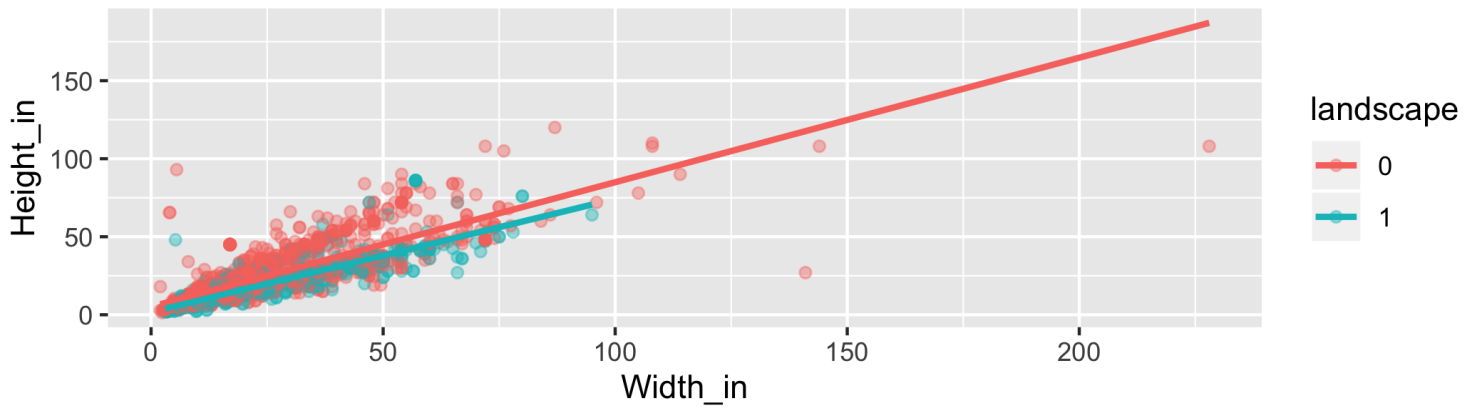
[1] Source: Wikipedia, Landscape painting

[2] Source: Wikipedia, Portait painting

Multiple explanatory variables

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

```
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(color = "landscape")
```



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 modelling 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

Want to follow along?

Go to RStudio Cloud -> make a copy of "Modeling Paris Paintings"

Height & width

```
m_ht_wt <- lm(Height_in ~ Width_in, data = pp)
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.
  - Does this make sense?

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/

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 the sample statistics to estimate them:

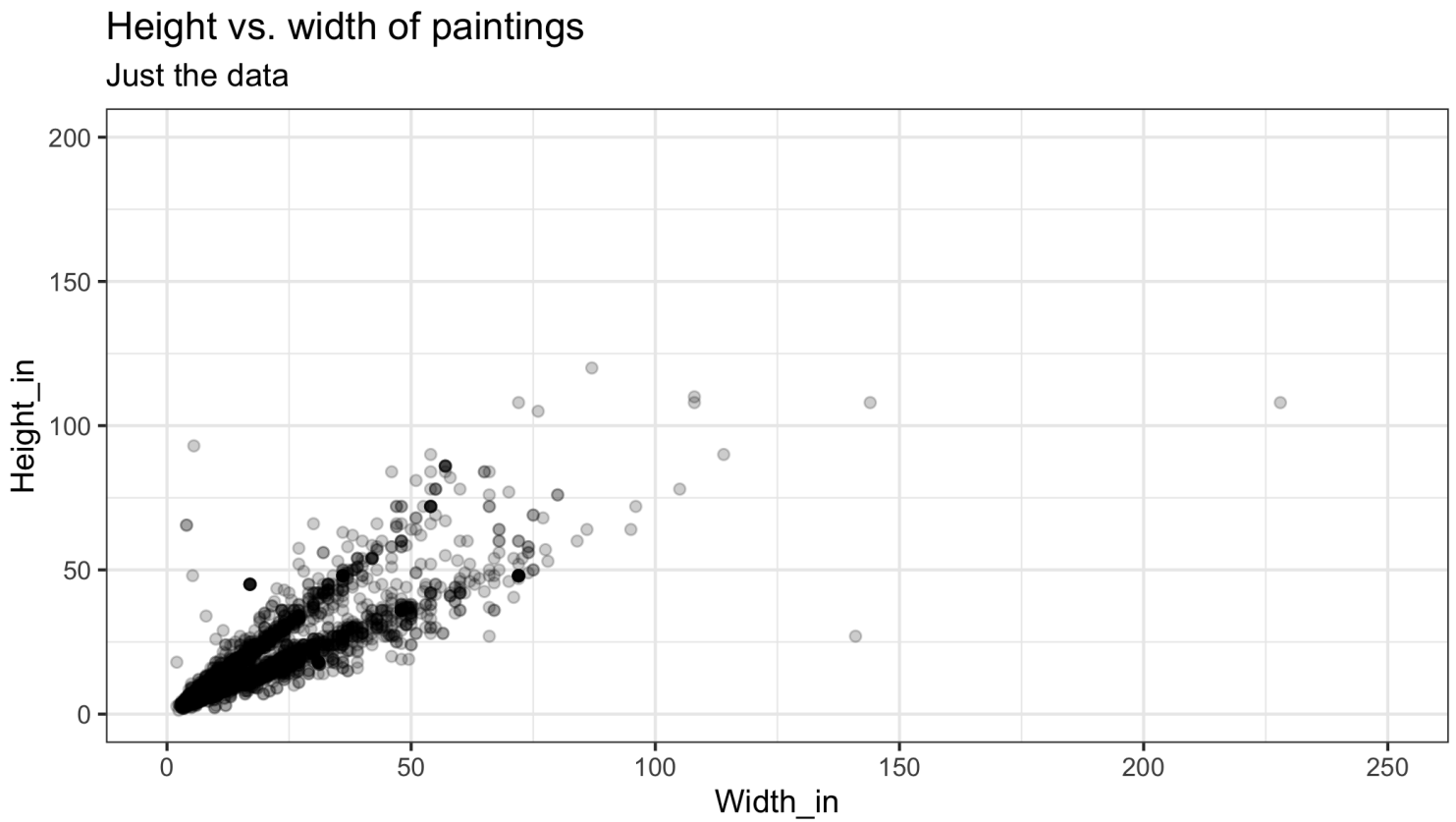
$$\hat{y} = b_0 + b_1 \, x$$

Least squares regression

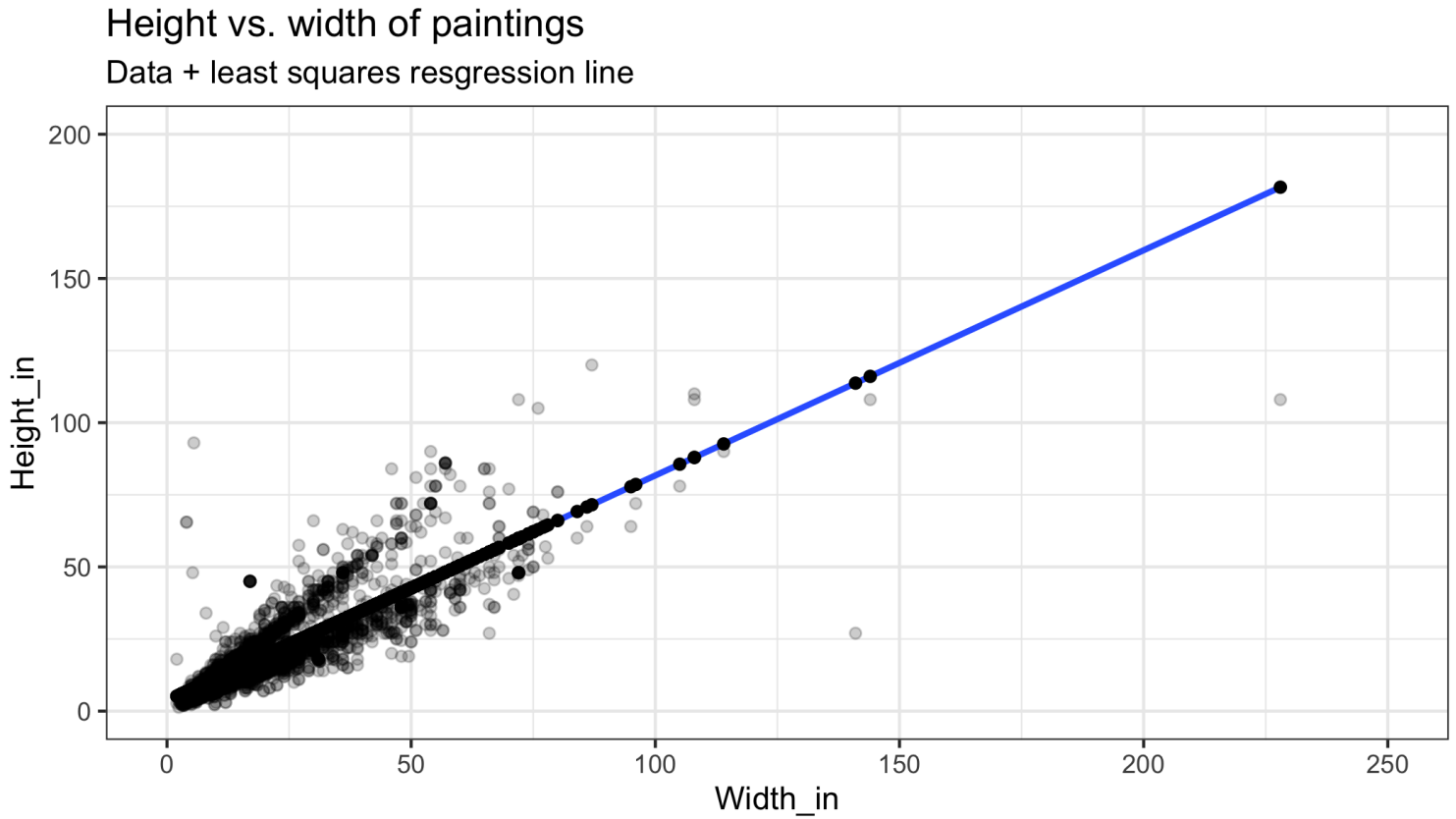
The regression line minimizes the sum of squared residuals.

- Residuals: $e_i = y - \hat{y}$,
- The regression line minimizes $\sum_{i=1}^{n} e_i^2$.

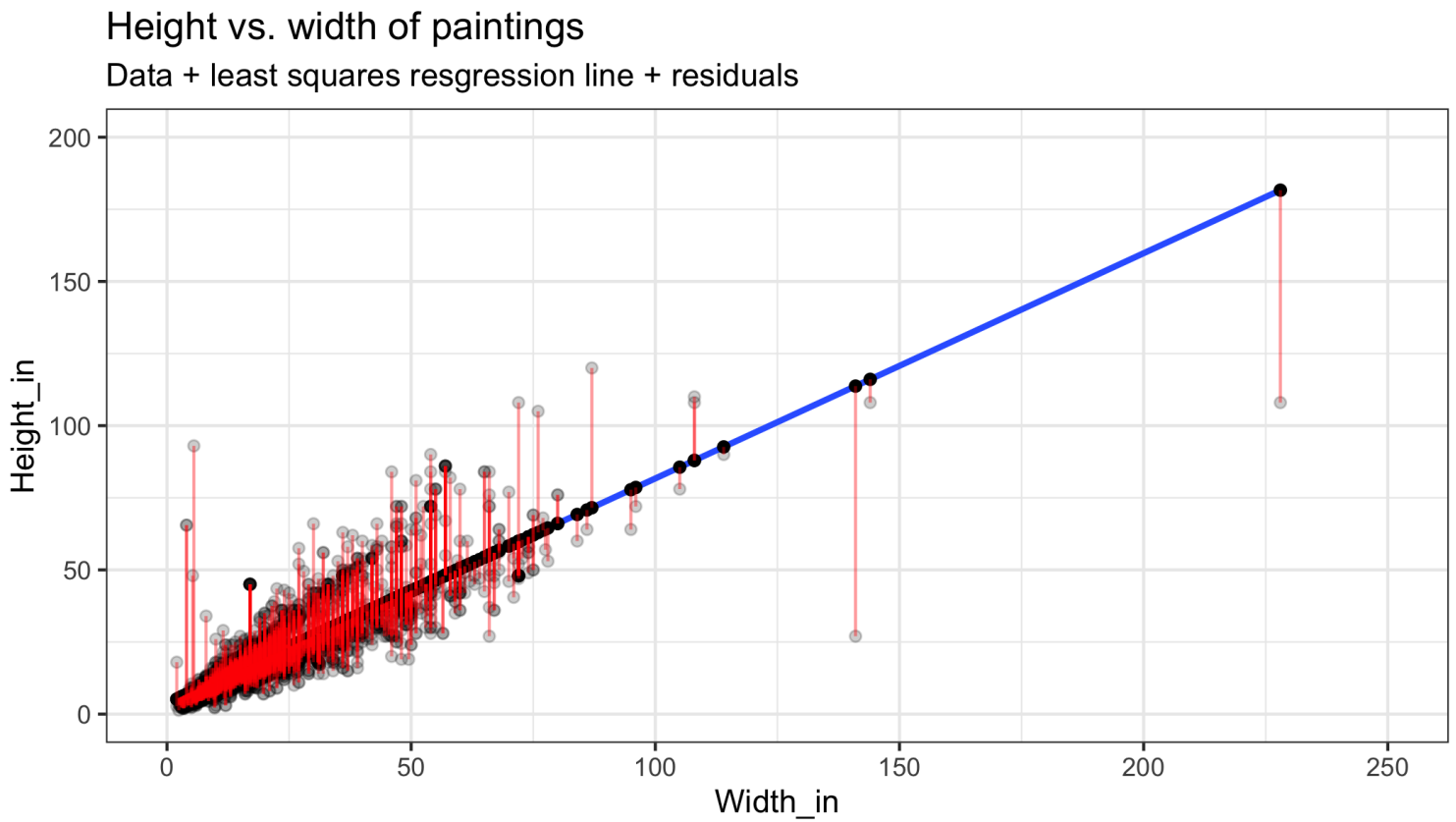
Visualizing residuals



Visualizing residuals (cont.)



Visualizing residuals (cont.)



Properties of the least squares regression line

- The slope has the same sign as the correlation coefficient:

$$b_1 = r\frac{s_y}{s_x}$$

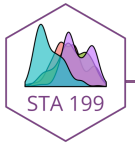
- The regression line goes through the center of mass point, the coordinates corresponding to average x and average y: $(\bar{x}, \bar{y})$.

$$\hat{y} = b_0 + b_1 x \quad \Rightarrow \quad b_0 = \bar{y} - b_1\bar{x}$$

- The sum of the residuals is zero:

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

- The residuals and x values are uncorrelated.



Height & landscape features

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

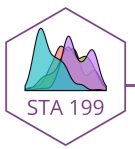
$$\widehat{Height_{in}} = 22.68 - 5.65\ landsALL$$

Height & landscape features (cont.)

- **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
```



Relationship between height and school

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

- 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.

Categorical predictor with >2 levels

```
## # 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
```

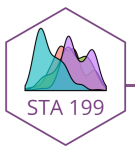
The linear model with multiple predictors

- Population model:

$$\hat{y} = \beta_0 + \beta_1 \, x_1 + \beta_2 \, x_2 + \cdots + \beta_k \, x_k$$

- Sample model that we use to estimate the population model:

$$\hat{y} = b_0 + b_1 \, x_1 + b_2 \, x_2 + \cdots + b_k \, x_k$$



Correlation does not imply causation!

Remember this when interpreting model coefficients