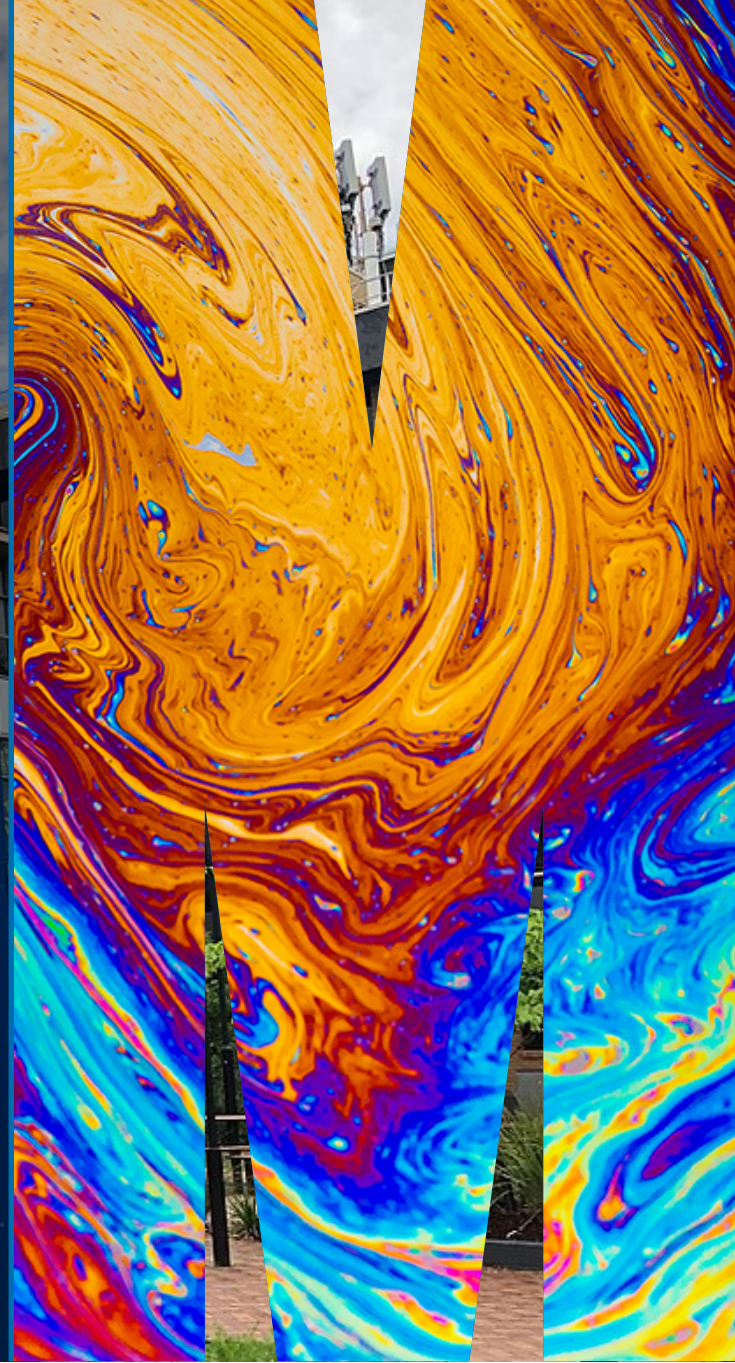
ETC5510: Introduction to Data Analysis

Week 7, part A

Linear Models



Lecturer: *Nicholas Tierney & Stuart Lee*
Department of Econometrics and Business Statistics
nicholas.tierney@monash.edu
May 2020

Midsemester test

- Thanks for completing on time!
- Note that final grade reported by Moodle is not your final grade!
- Will put together the final grades over the next week!

Assignment 2 notes

- Will be posted on the website on Monday.
- Some questions might seem a bit strange - this is normal!
- Sometimes you can't answer the exact question with the data you have.
- Steer away from the idea of "correct code" and reframe this as "code that works". There is no one single answer - remember the Tower of Babel example from the start of class, there are many ways to do the same thing in R.

Assignment 2 notes

- Our marking advice: Describe how you think the solution you have provided helps answer the question.
  - Show us a plot
  - Explain what you see in the plot - what do you think we can learn from the information we have?
  - Describe what other information you might like to have

Recap

- style
- functions
- map

Today: The language of models

Modelling

- 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

- You're familiar with the tidyverse:
- The broom package takes the messy output of built-in functions in R, such as `lm`, and turns them into tidy data frames.

Data: Paris Paintings

Paris Paintings

```
pp <- read_csv(here::here("slides/data/paris-paintings.csv"),
               na = c("n/a", "", "NA"))
pp
## # A tibble: 3,393 x 61
##    name  sale  lot   position dealer  year origin_author origin_cat school_pntg
##    <chr> <chr> <chr>    <dbl> <chr>  <dbl> <chr>         <chr>      <chr>
##  1 L176… L1764 2       0.0328 L       1764 F             0          F
##  2 L176… L1764 3       0.0492 L       1764 I             0          I
##  3 L176… L1764 4       0.0656 L       1764 X             0          D/FL
##  4 L176… L1764 5       0.0820 L       1764 F             0          F
##  5 L176… L1764 5       0.0820 L       1764 F             0          F
##  6 L176… L1764 6       0.0984 L       1764 X             0          I
##  7 L176… L1764 7       0.115  L       1764 F             0          F
##  8 L176… L1764 7       0.115  L       1764 F             0          F
##  9 L176… L1764 8       0.131  L       1764 X             0          I
## 10 L176… L1764 9       0.148  L       1764 D/FL          0          D/FL
## # … with 3,383 more rows, and 52 more variables: diff_origin <dbl>,
## #   logprice <dbl>, price <dbl>, count <dbl>, subject <chr>,
## #   authorstandard <chr>, artistliving <dbl>, authorstyle <chr>, author <chr>,
## #   winningbidder <chr>, winningbiddertype <chr>, endbuyer <chr>, Interm <dbl>,
```

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

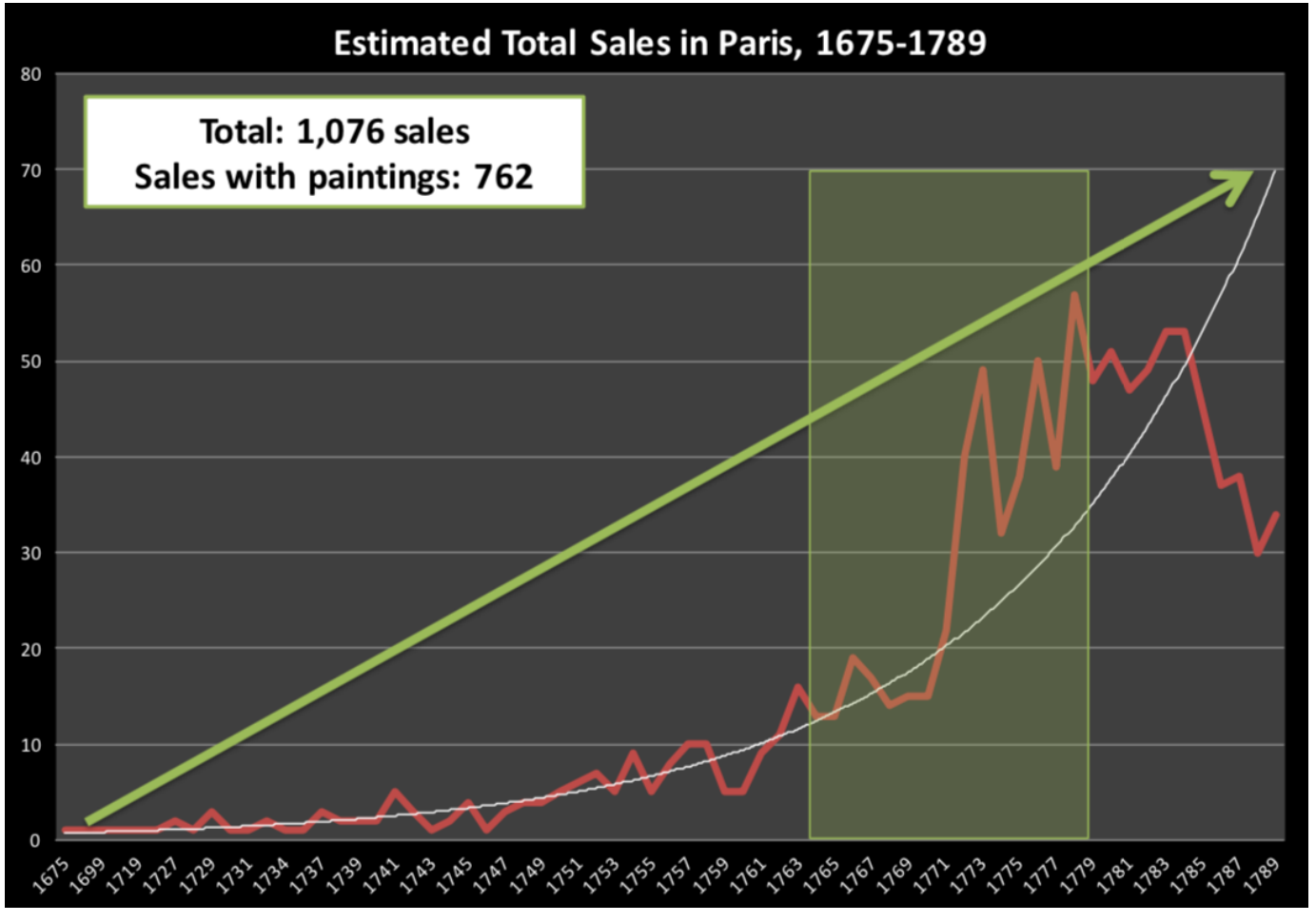
Auctions today

Old Master & British Paintings Evening Sale Soars over Estimate

Auctions back in the day

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

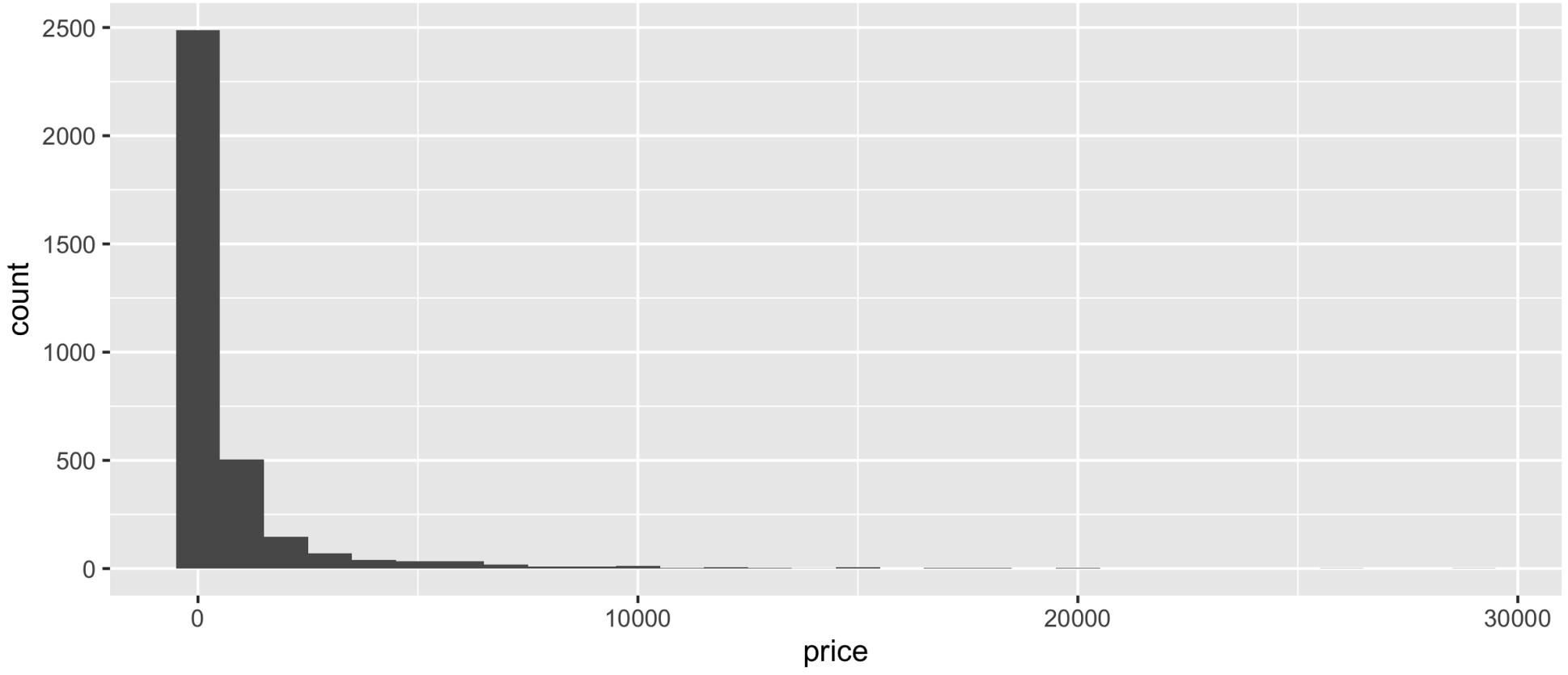
Paris auction market



Modelling the relationship between variables

Prices: Describe the distribution of prices of paintings.

```
ggplot(data = pp, aes(x = price)) +
  geom_histogram(binwidth = 1000)
```



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

Models as functions: Example

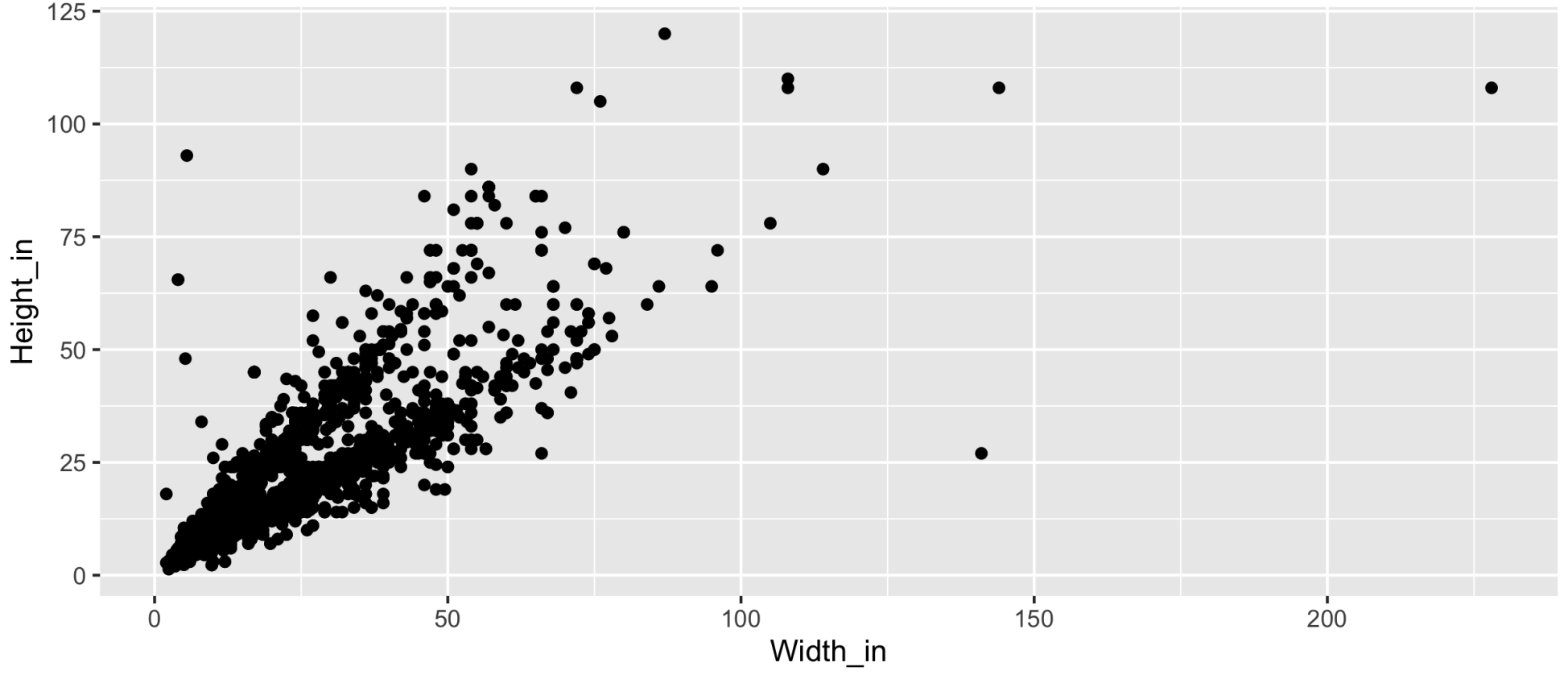
- The formula $y = 3x + 7$ is a function with input x and output y, when x is 5, the output y is 22

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

```
anon <- function(x) 3*x + 7
anon(5)
## [1] 22
```

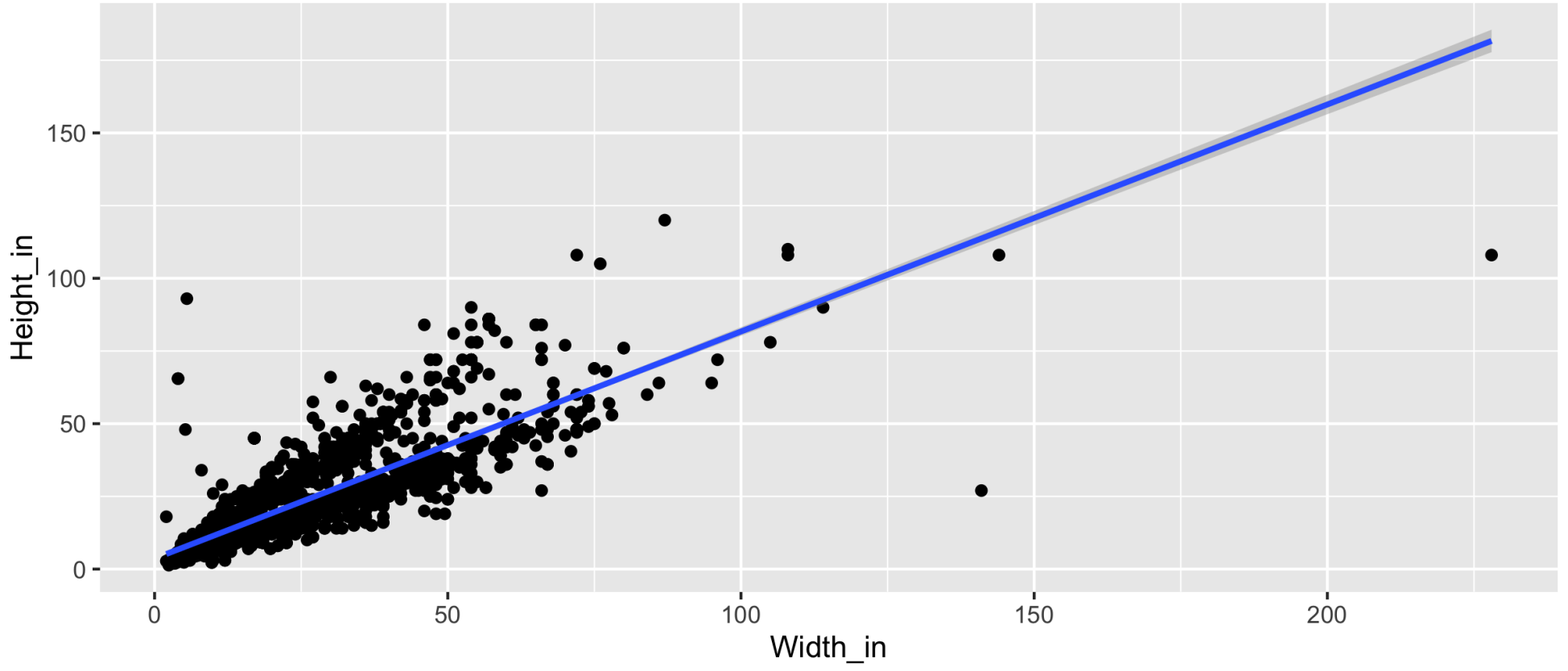
Height as a function of width

Describe the relationship between height and width of paintings.



Visualizing the linear model

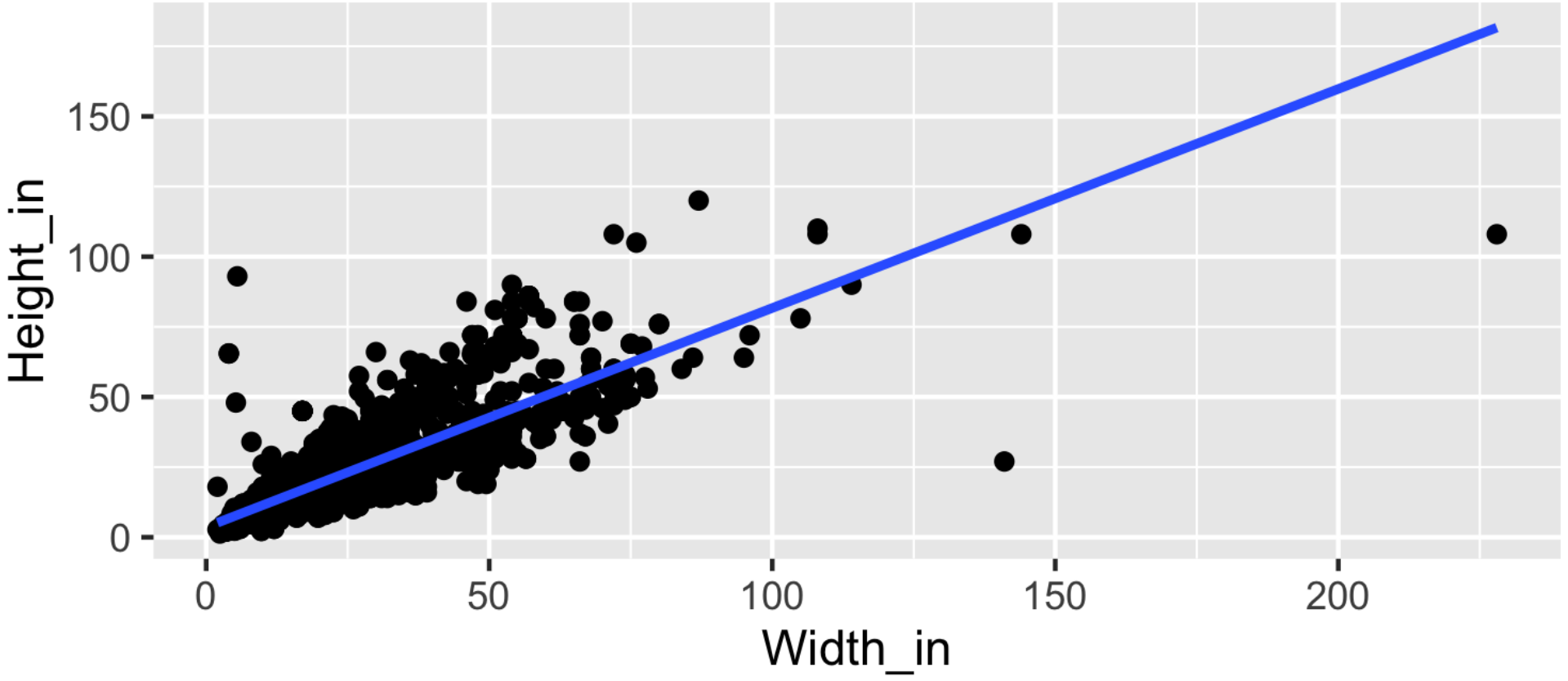
```
ggplot(data = pp, aes(x = Width_in, y = Height_in)) +
  geom_point() +
  geom_smooth(method = "lm") # lm for linear model
```



Visualizing the linear model

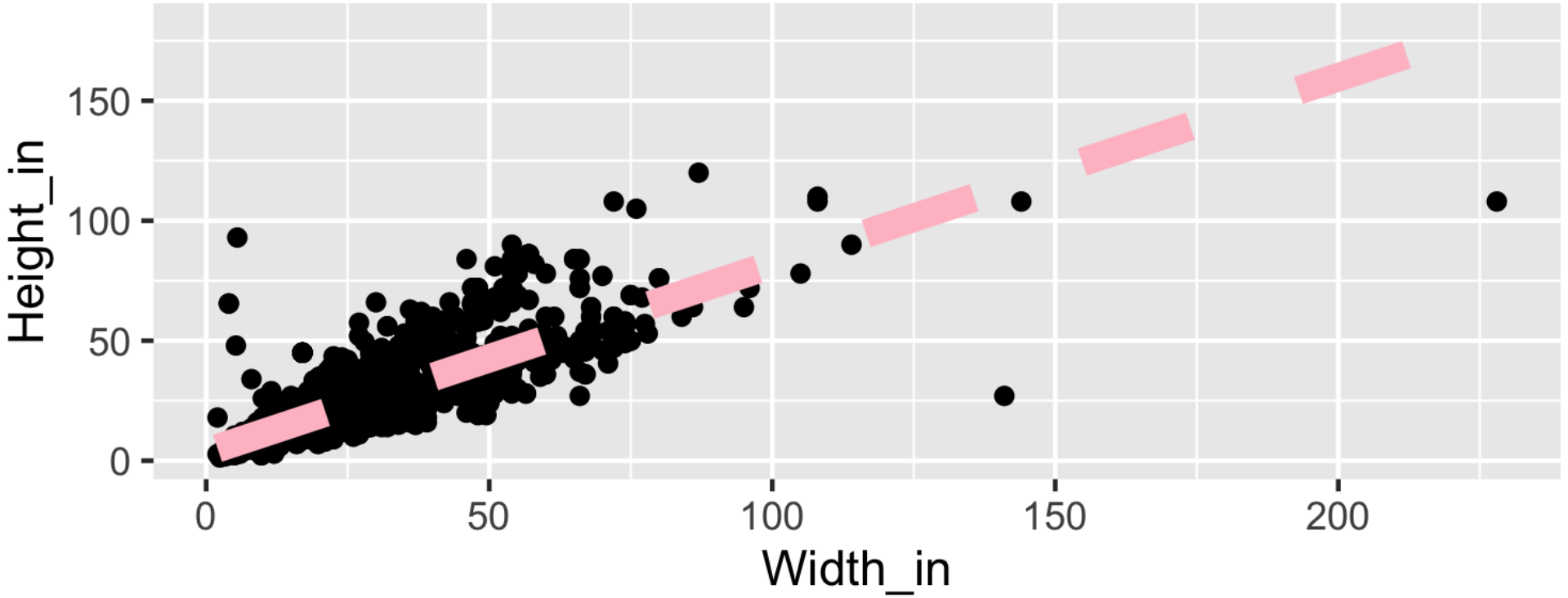
(without the measure of uncertainty around the line)

```
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 (style the line)

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



Vocabulary

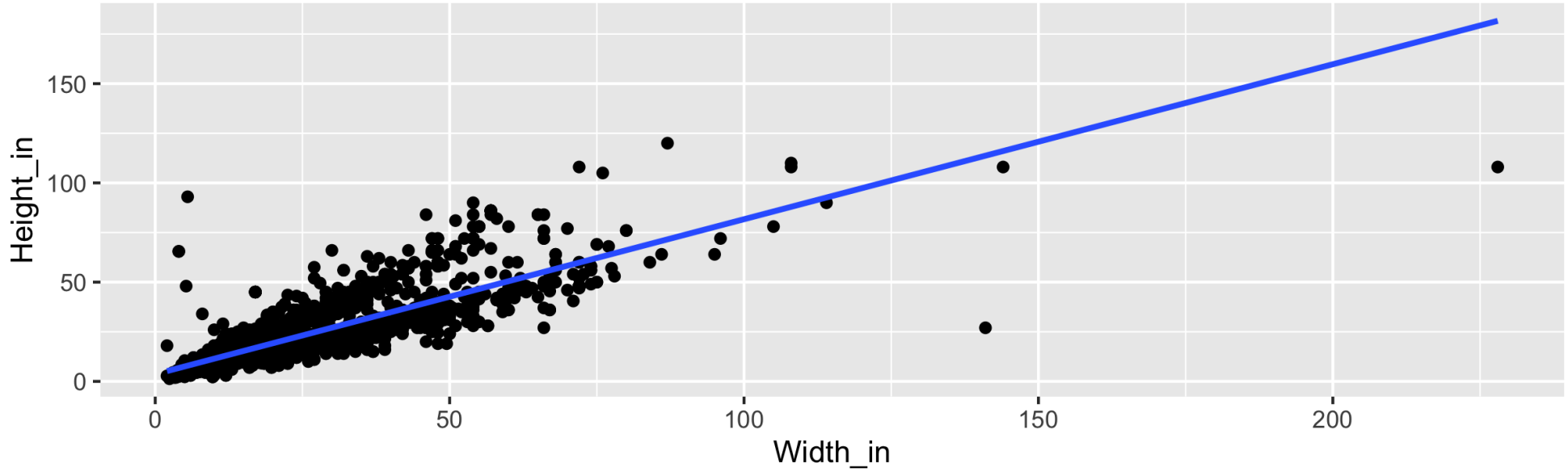
- **Response variable:** Variable whose behavior or variation you are trying to understand, on the y-axis (dependent variable)
- **Explanatory variables:** Other variables that you want to use to explain the variation in the response, on the x-axis (independent variables)

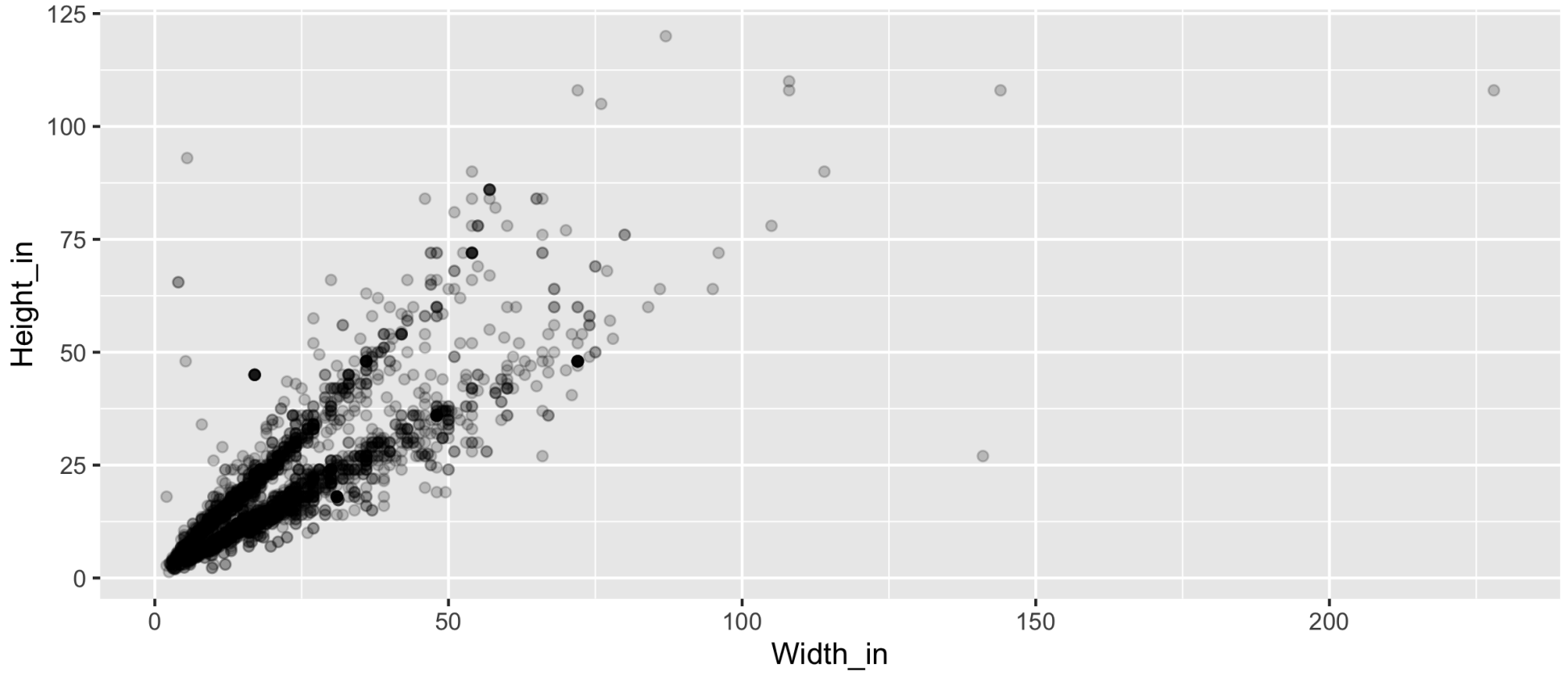
Vocabulary

- **Predicted value:** Output of the function **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?





- 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 vs portait 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.

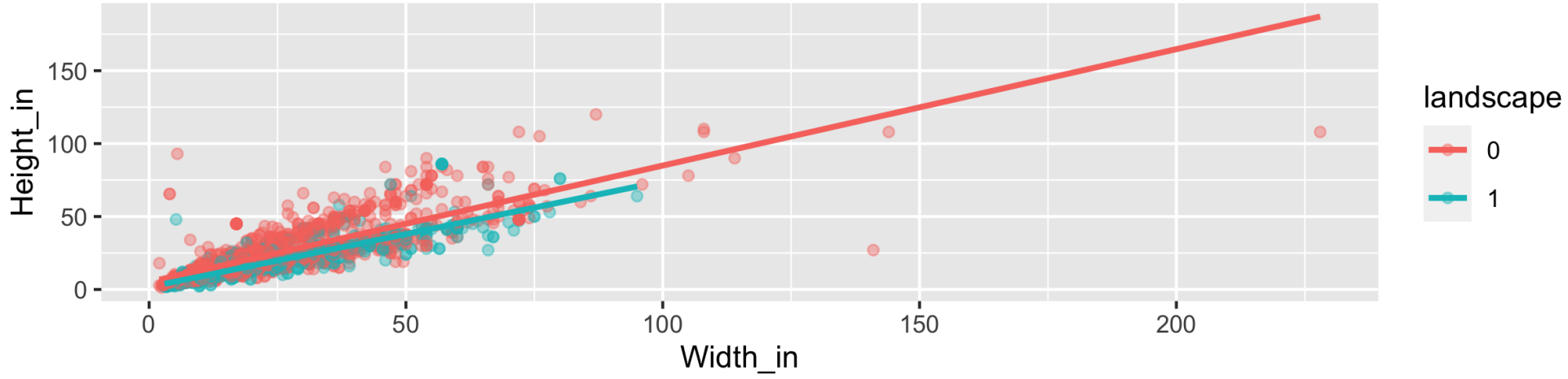
[1] Source: Wikipedia, Landscape painting

[2] Source: Wikipedia, Portait painting

Multiple explanatory variables

How, if at all, the relatonship 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.

- Scatterplot suggests 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. (also called **inference**, as you **make inference** on these relationships)
2. Prediction: Plug in x, get the predicted y

Your Turn: go to Rstudio and begin the excercise 7a

```
library(countdown)
countdown(minutes = 6)
```

06:00

Characterizing relationships with models

Height & width

```
m_ht_wt <- lm(Height_in ~ Width_in, data = pp)
m_ht_wt
##
## Call:
## lm(formula = Height_in ~ Width_in, data = pp)
##
## Coefficients:
## (Intercept)     Width_in
##      3.6214       0.7808
```

Model of height and width

$$\widehat{Height}_{in} = 3.62 + 0.78\ Width_{in}$$

- **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

- Interested in β_0 (population parameter for the intercept) and the β_1 (population parameter for the slope) in the following model:

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

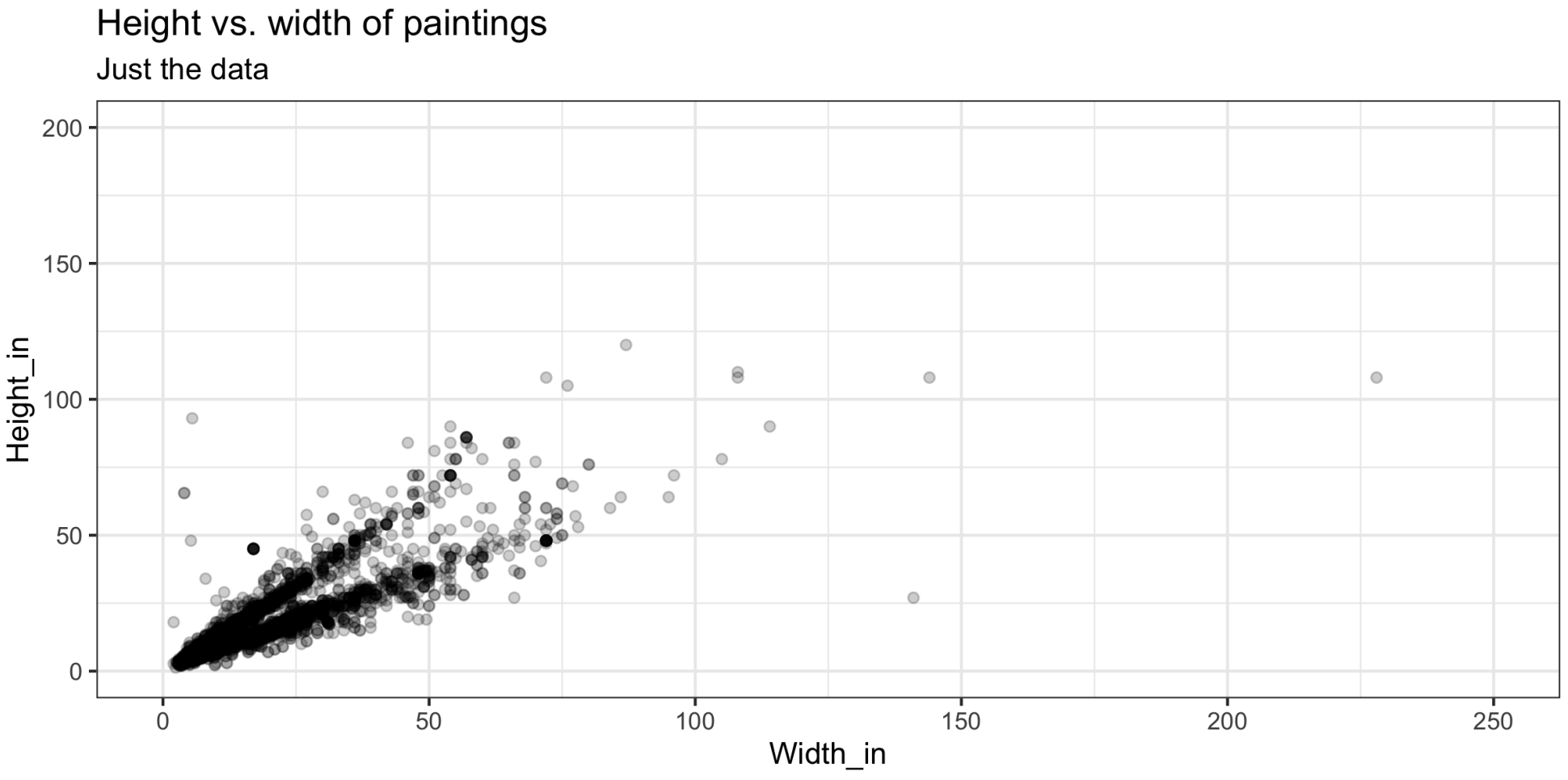
Least squares regression

The regression line minimizes the sum of squared residuals.

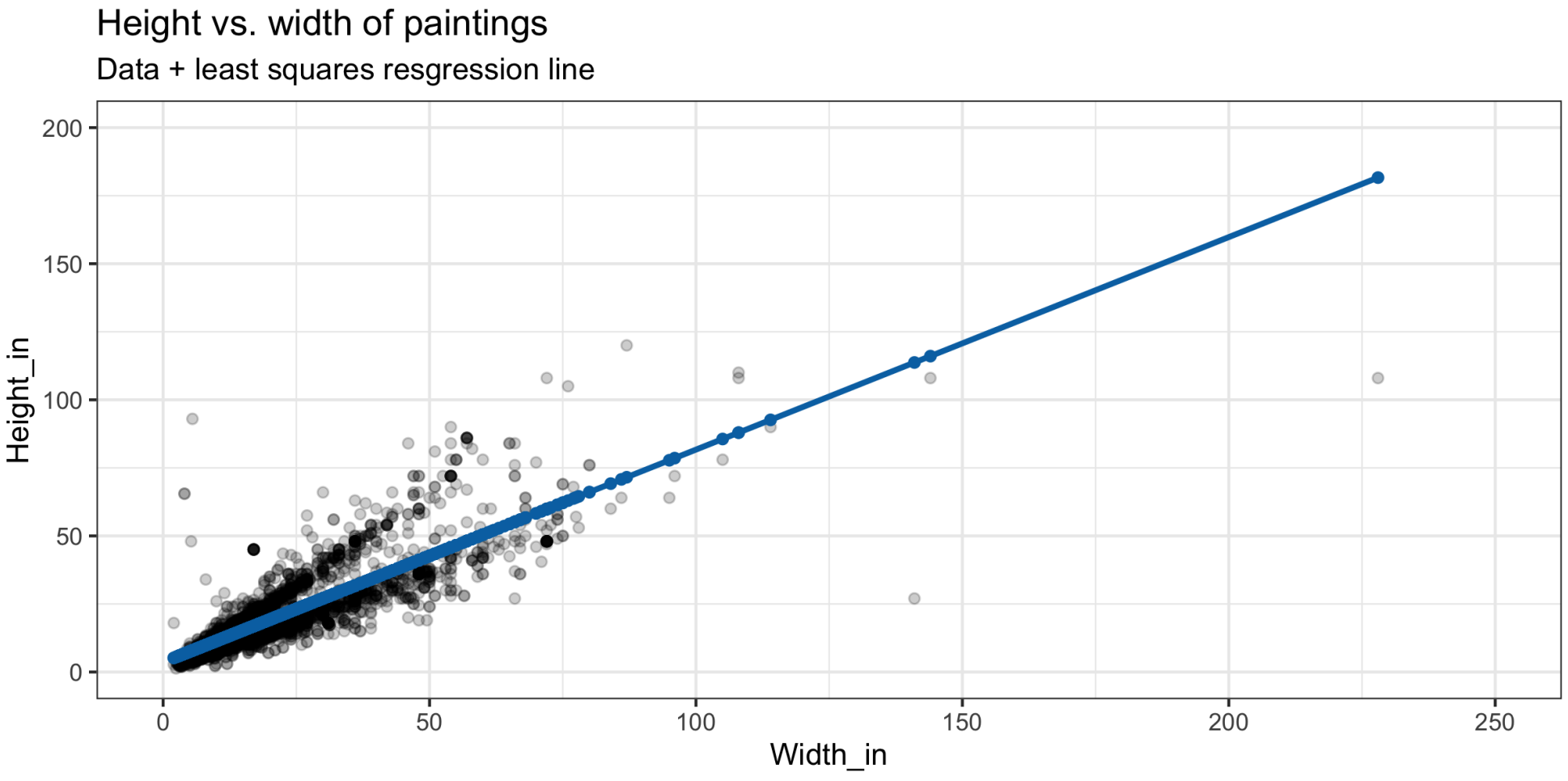
If $e_i = y - \hat{y}$,

then, the regression line minimizes $\sum_{i=1}^{n} e_i^2$.

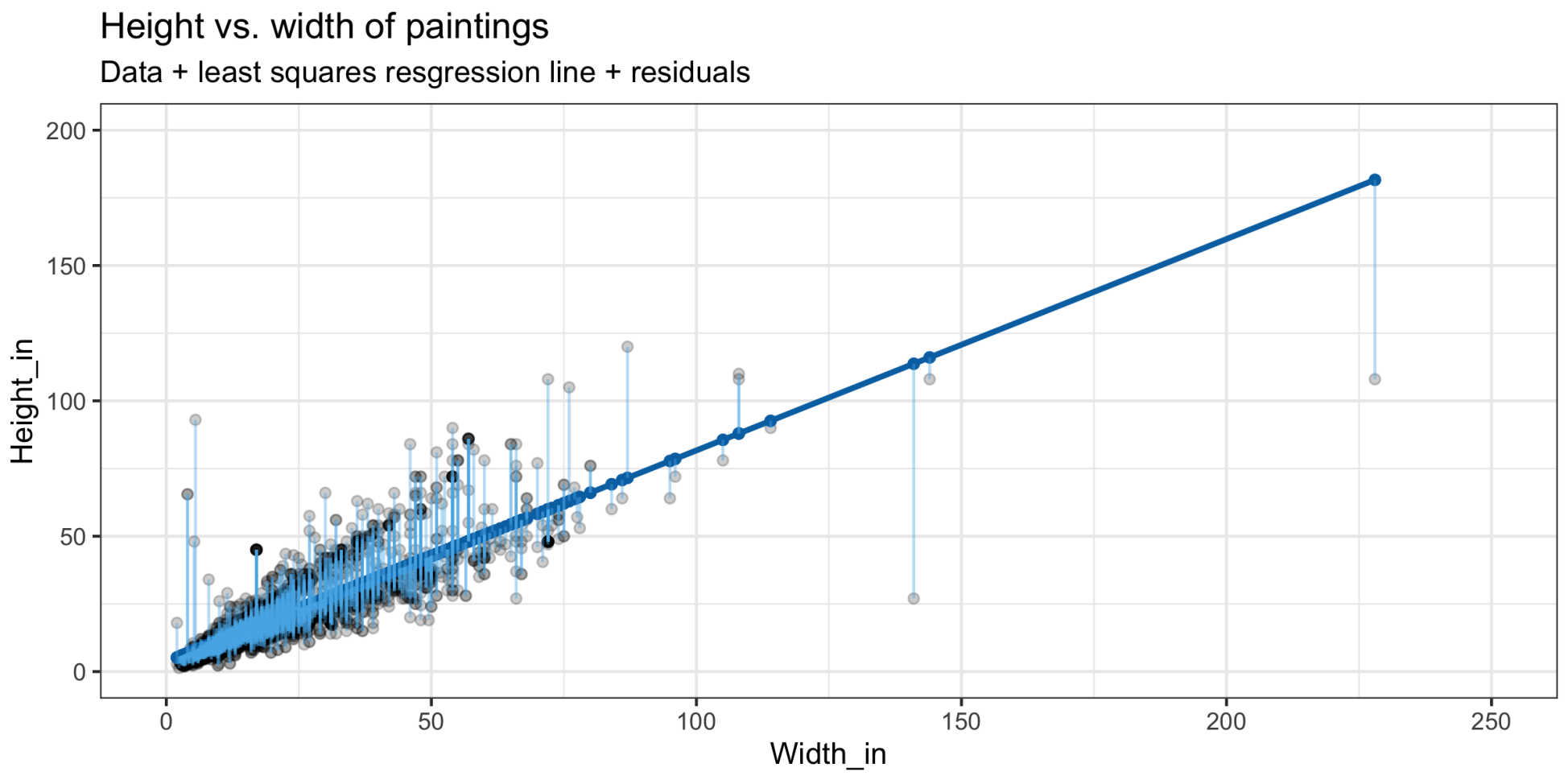
Visualizing residuals



Visualizing residuals (cont.)



Visualizing residuals (cont.)



Properties of the least squares regression line

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

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

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

$$\beta_1 = r\frac{s_y}{s_x}$$

Assumptions of least squares regression line

- The sum of the residuals is zero:

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

- The residuals have constant variance (homoskedastic errors)
- The residuals and x values are uncorrelated.

Height & landscape features

```
m_ht_lands <- lm(Height_in ~ factor(landsALL), data = pp)
m_ht_lands
##
## Call:
## lm(formula = Height_in ~ factor(landsALL), data = pp)
##
## Coefficients:
##       (Intercept)  factor(landsALL)1
##            22.680             -5.645
```

$$\widehat{Height}_{in} = 22.68 - 5.65\ landsALL$$

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))
##
## Call:
## lm(formula = Height_in ~ school_pntg, data = pp)
##
## Coefficients:
##     (Intercept)  school_pntgD/FL     school_pntgF     school_pntgG
##          14.000            2.329           10.197            1.650
##    school_pntgI     school_pntgS     school_pntgX
##          10.287           30.429            2.869
```

- When the categorical explanatory variable has many levels, they're 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

Prediction with models

Predict height from width

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

$$\widehat{Height}_{in} = 3.62 + 0.78\ Width_{in}$$

```
3.62 + 0.78 * 60
## [1] 50.42
```

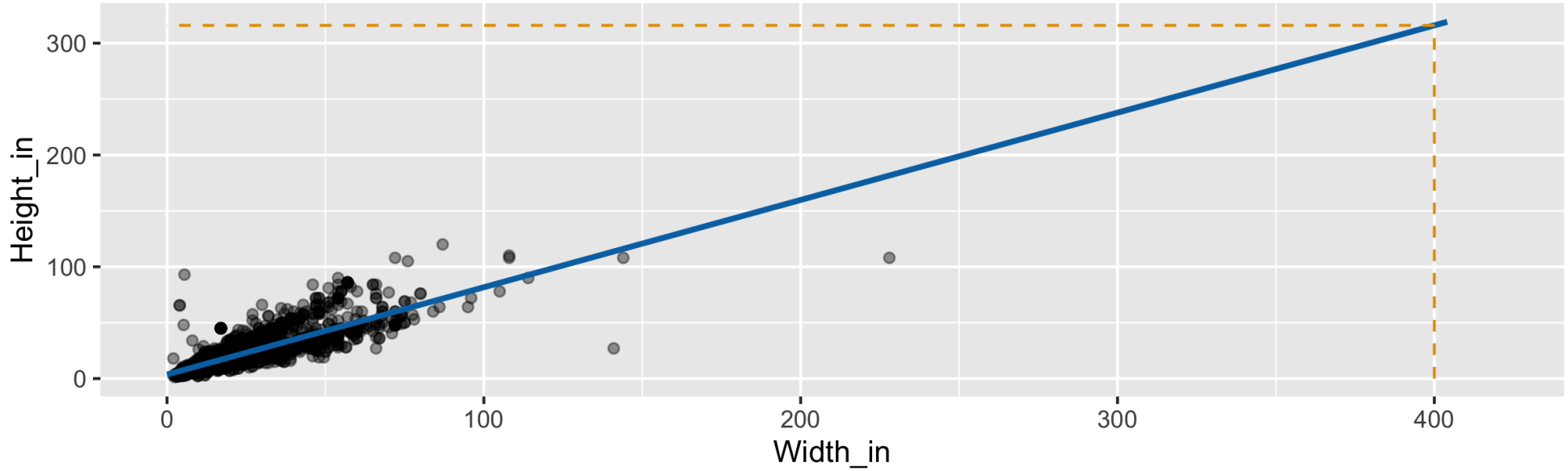
"On average, we expect paintings that are 60 inches wide to be 50.42 inches high."

Warning: We "expect" this to happen, but there will be some variability. (We'll learn about measuring the variability around the prediction later.)

Prediction vs. extrapolation

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

$$\widehat{Height}_{in} = 3.62 + 0.78\ Width_{in}$$



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]*

(OpenIntro Statistics. "Extrapolation is treacherous." OpenIntro Statistics.)

Measuring model fit

Measuring the strength of the fit

- R^2 is a common measurement of strength of linear model fit.
- R^2 tells us % variability in response explained by model.
- Remaining variation is explained by variables not in the model.
- R^2 is sometimes called the coefficient of determination.

Obtaining R^2 in R

- Height vs. width

```
broom::glance(m_ht_wt)
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic p.value    df  logLik    AIC    BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>   <dbl>  <dbl>  <dbl>
## 1     0.683         0.683  8.30     6749.       0     2 -11083. 22173. 22191.
## # … with 2 more variables: deviance <dbl>, df.residual <int>
broom::glance(m_ht_wt)$r.squared # extract R-squared
## [1] 0.6829468
```

Roughly 68% of the variability in heights of paintings can be explained by their widths.

Obtaining R^2 in R

- Height vs. lanscape features

```
glance(m_ht_lands)$r.squared
## [1] 0.03456724
```

Your Turn: Go to Rstudio and complete the lab exercise

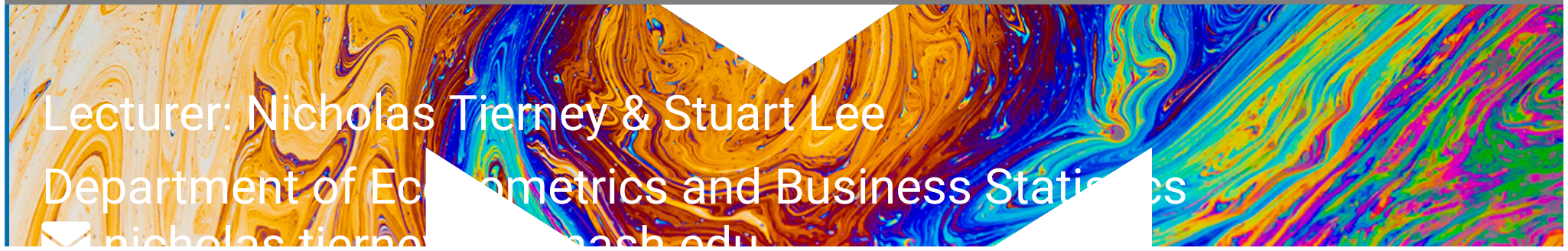
References

- data science in a box

background-image: url(images/bg1.jpg) background-size: cover
class: hide-slide-number split-70 count: false

That's it!

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.



Lecturer: Nicholas Tierney & Stuart Lee

Department of Ec metrics and Business Stati cs

nicholas.tierne ash.edu