Simple Regression Models

Introduction

Prof. Dr. Jan Kirenz

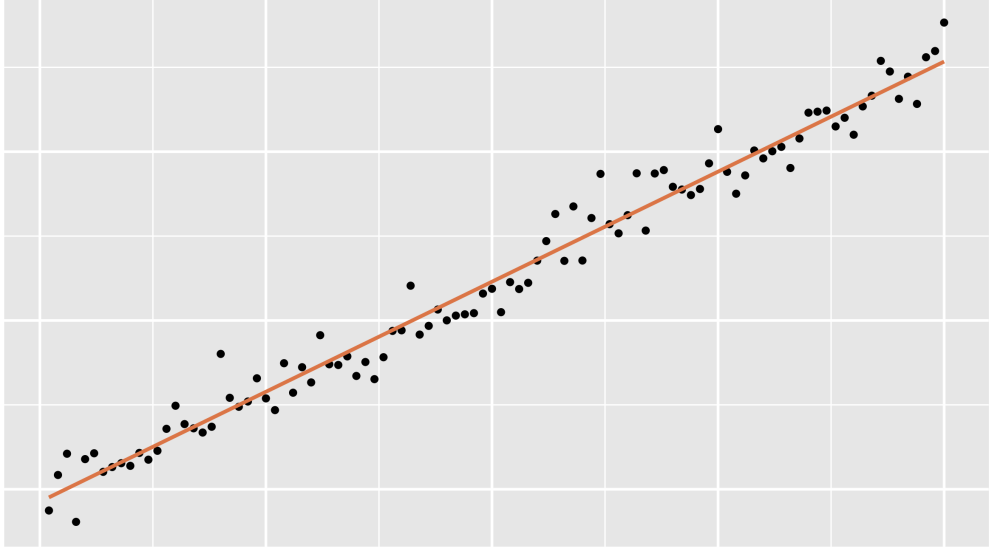
The following content is based on Mine Çetinkaya-Rundel's excellent book Data Science in a Box

What is a model?

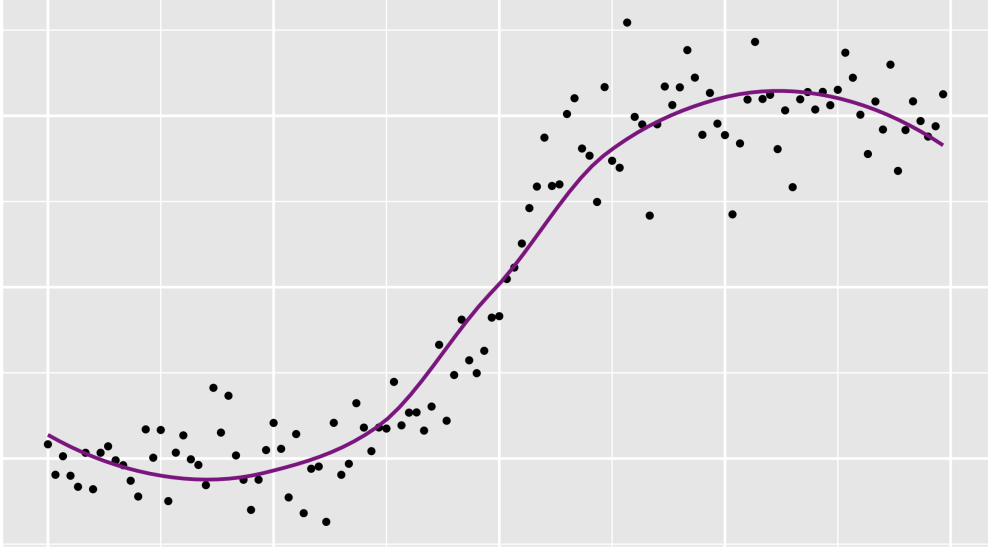
Modelling

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

Linear



Non-linear



Data: Paris Paintings

Paris Paintings

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

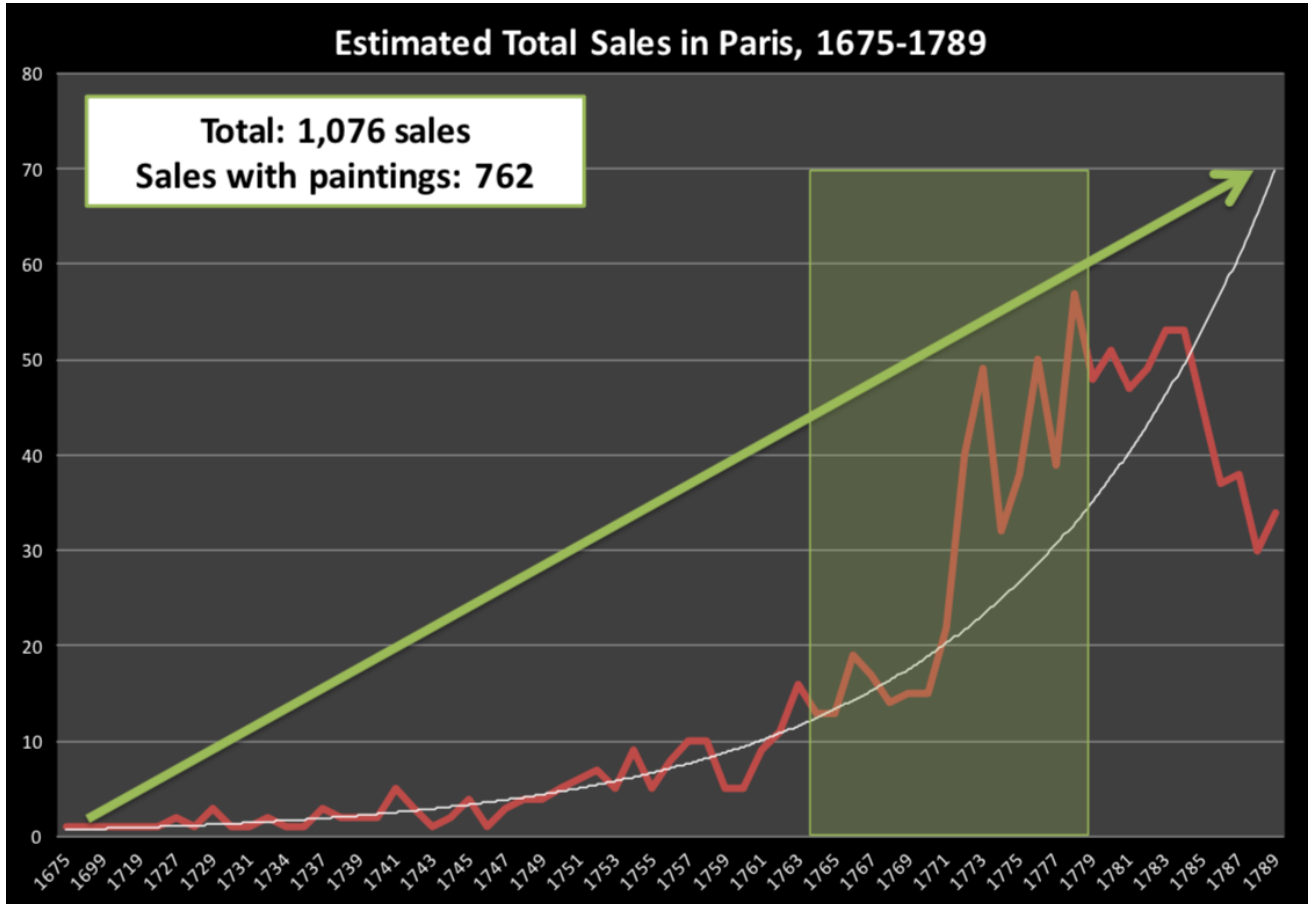
- Source: Printed catalogues of 28 auction sales in Paris, 1764 - 1780
- Data curators Sandra van Ginhoven and Hilary Coe Cronheim (who were PhD students in the Duke Art, Law, and Markets Initiative at the time of putting together this dataset) translated and tabulated the catalogues
- 3393 paintings, their prices, and descriptive details from sales catalogues over 60 variables

Auctions today

Auctions back in the day

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

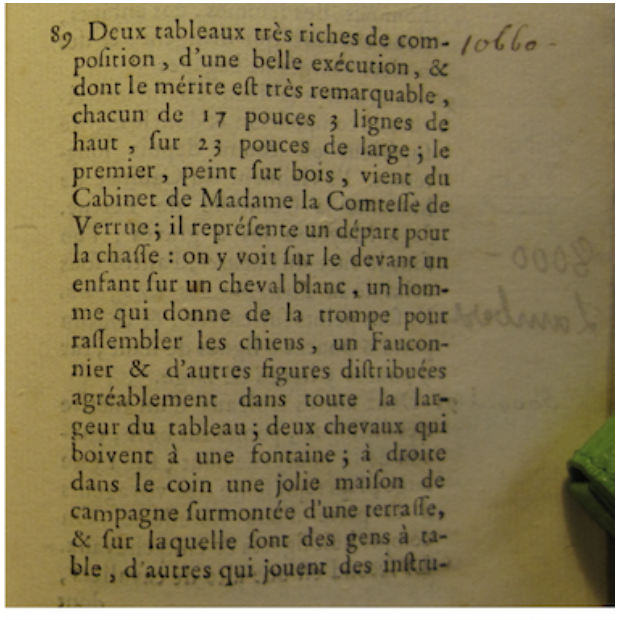
Paris auction market



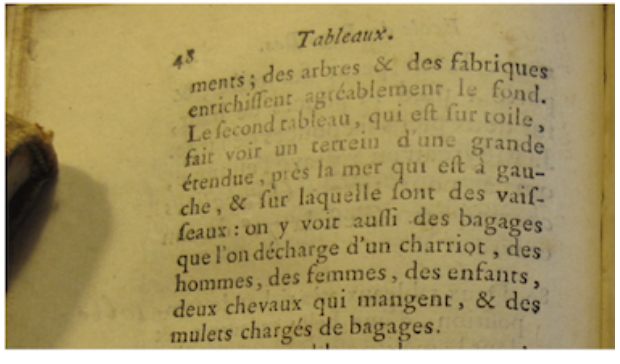
Plot credit: Sandra van Ginhoven

Départ pour la chasse

Auction catalog text



89 Deux tableaux très riches de compoſition, 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 homme qui donne de la trompe pour raſſembler les chiens, un Fauconnier & d'autres figures diſtribuées agréablement dans toute la largeur 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 table, 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 gauche, & ſur laquelle ſont des vaiſſeaux: on y voit auſſi des bagages que l'on décharge d'un charriot, des hommes, des femmes, des enfants, deux chevaux qui mangent, & des mulets chargés de bagages.

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.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
1	name	sale	lot	dealer	year	origin_author	origin_cat	school_pntg	diff_origin	price	count	subject	authorstandard	artistliving	authorstyle	author	winnir
2517	R1777-86	R1777	86	R	1777	D/FL	D/FL	D/FL	0	620.0	1	2 femmes, enfants, paysage vu à travers une arcade	Bega, Cornelis Pieterszoon	0	n/a	Corneille Bega	Lebrur
2518	R1777-87	R1777	87	R	1777	D/FL	D/FL	D/FL	0	12,000.0	1	Course du hareng	Wouwerman, Philips	0	n/a	Philippe Wouwerman	Donjeu
2519	R1777-88	R1777	88	R	1777	D/FL	D/FL	D/FL	0	8,000.0	1	Paysage sablonneux	Wouwerman, Philips	0	n/a	Philippe Wouwerman	Lambe
2520	R1777-89a	R1777	89	R	1777	D/FL	D/FL	D/FL	0	5,300.0	1	Départ pour la chasse	Wouwerman, Philips	0	n/a	Philippe Wouwerman	Langlie
2521	R1777-89b	R1777	89	R	1777	D/FL	D/FL	D/FL	0	5,300.0	1	Déchargement d'un chariot, estran	Wouwerman, Philips	0	n/a	Philippe Wouwerman	Langlie

	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
1	winningbidder	winningbiddertype	endbuyer	Interm	type_interm ed	Height_in	Width_in	Surface_Rect	Diam_in	Surface_Rnd	Shape	Surface	material	mat	quantity	nfigures	engraved
2516	Feuillet	D	D	0		16	20	320			squ_rect	320	toile	t	1	0	0
2517	Lebrun, Jean-Baptiste-Pierre	D	D	0		13.25	11	145.75			squ_rect	145.75	bois	b	1	0	0
2518	Donjeux, Vincent	D	D	0		23	29.25	672.75			squ_rect	672.75	toile	t	1	50	0
2519	Lambert, John (Chevalier Lambert)	C	C	0		23	30	690			squ_rect	690	toile	t	1	0	1
2520	Langlier, Jacques for Poullain, Antoine	DC	C	1	D	17.25	23	396.75			squ_rect	396.75	bois	b	1	0	0
2521	Langlier, Jacques for Poullain, Antoine	DC	C	1	D	17.25	23	396.75			squ_rect	396.75	toile	t	1	0	0
2522	Choiseul-Praslin, Comte de	C	C	0		6.5	9.25	60.125			squ_rect	60.125	cuivre	c	1	0	0

	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF
1	nfigures	engraved	original	prevcoll	othartist	paired	figures	finished	lrgfont	relig	landsALL	lands_sc	lands_figs	lands_ment	arch	mytho	peasant	othgenre	singlefig	portrait	still_life	discauth	history	allegory	pastorale	other
2516	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2517	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
2518	50	0	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
2519	0	1	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2520	0	0	0	1	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2521	0	0	0	1	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2522	0	0	0	1	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

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

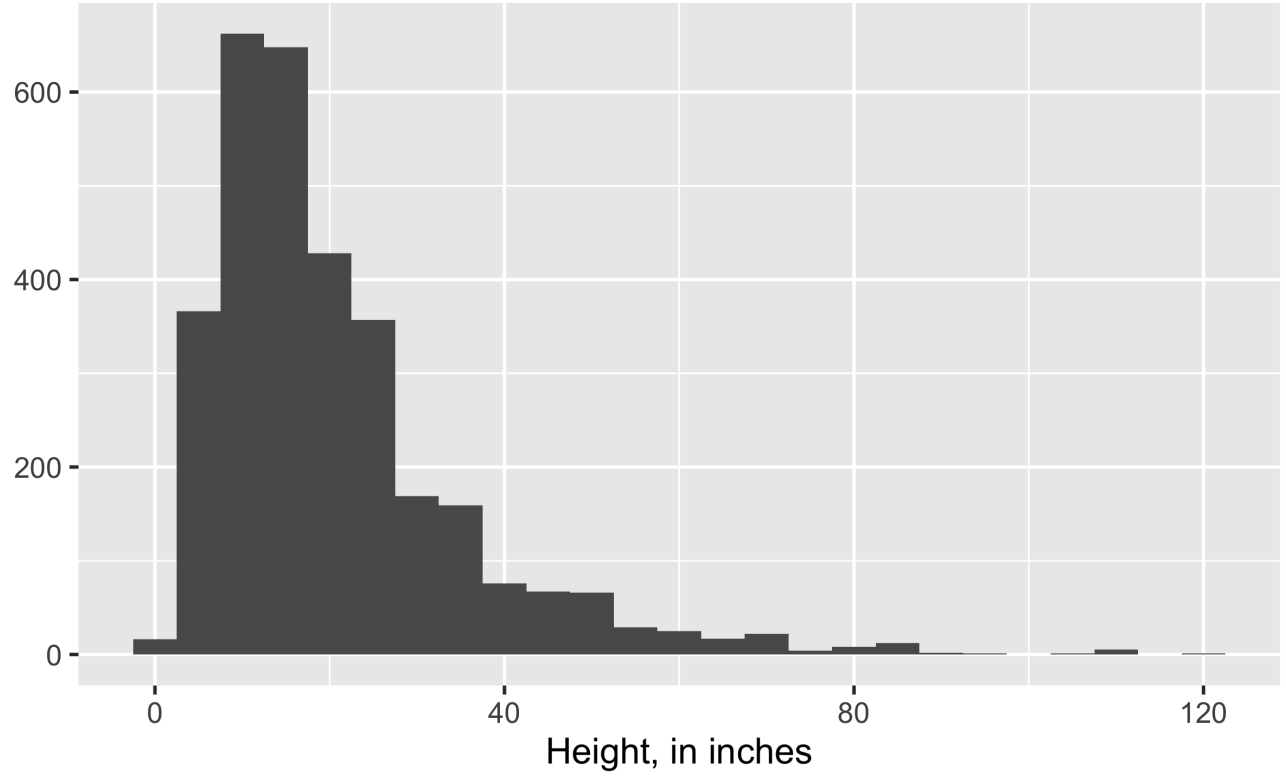
```
## Rows: 1
## Columns: 61
## $ name              <chr> "R1777-89a"
## $ sale              <chr> "R1777"
## $ lot               <chr> "89"
## $ position          <dbl> 0.3755274
## $ dealer            <chr> "R"
## $ year              <dbl> 1777
## $ origin_author     <chr> "D/FL"
## $ origin_cat        <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\u008epart pour la chasse"
## $ authorstandard    <chr> "Wouwerman, Philips"
## $ artistliving      <dbl> 0
## $ authorstyle       <chr> NA
## $ author            <chr> "Philippe Wouwermans"
## $ winningbidder     <chr> "Langlier, Jacques for Poullain, Ant…
## $ 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
...
```

Modeling the relationship between variables

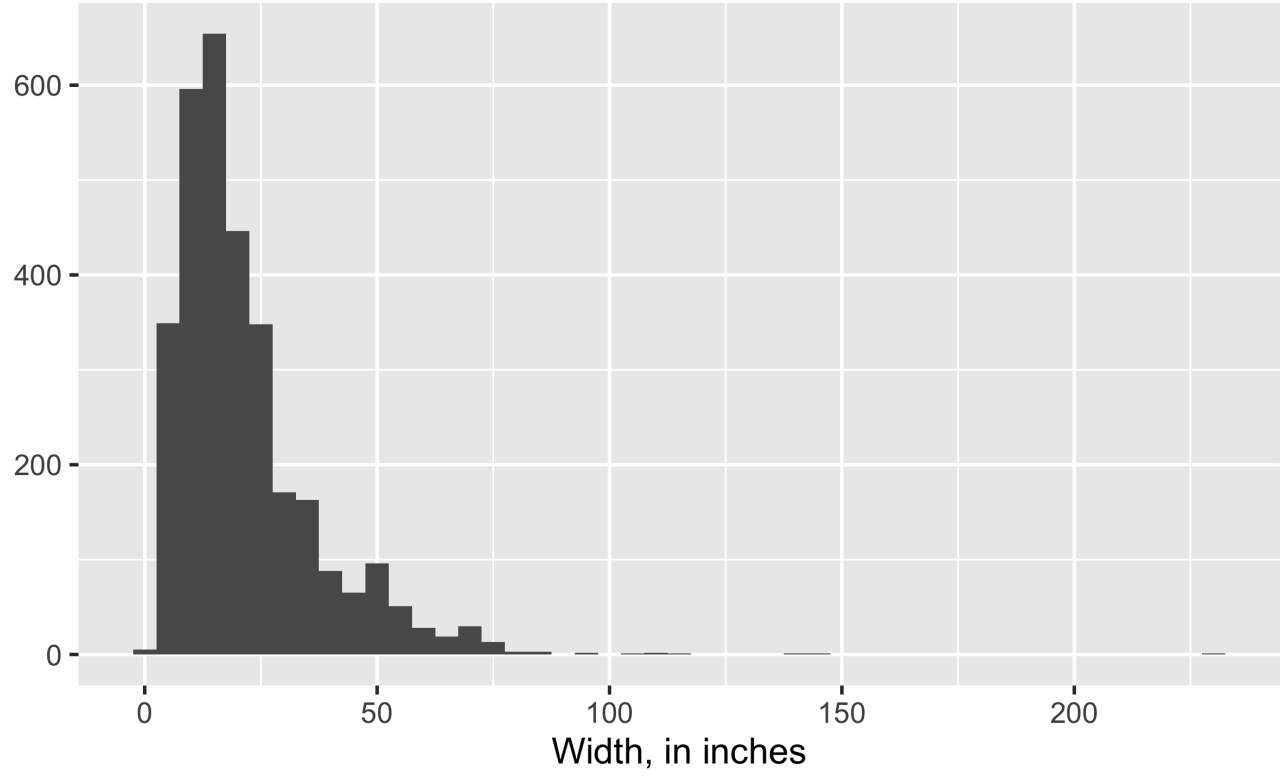
Heights

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



Widths

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



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.
- If x is 5, y is 22,

$$y = 3 \times 5 + 7 = 22$$

Height as a function of width

Plot Code

```
ggplot(data = pp, aes(x = Width_in, y = Height_in)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(
    title = "Height vs. width of paintings",
    subtitle = "Paris auctions, 1764 - 1780",
    x = "Width (inches)",
    y = "Height (inches)"
  )
```

```
## Warning: Removed 258 rows containing non-finite values
## (stat_smooth).

## Warning: Removed 258 rows containing missing values (geom_point).
```

... without the measure of uncertainty

Plot Code

```
ggplot(data = pp, aes(x = Width_in, y = Height_in)) +
  geom_point() +
  geom_smooth(method = "lm",
              se = FALSE) +
  labs(
    title = "Height vs. width of paintings",
    subtitle = "Paris auctions, 1764 - 1780",
    x = "Width (inches)",
    y = "Height (inches)"
  )
```

... with different cosmetic choices

Plot Code

```
ggplot(data = pp, aes(x = Width_in, y = Height_in)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE,
              color = "#8E2C90", linetype = "dashed", size = 3) +
  labs(
    title = "Height vs. width of paintings",
    subtitle = "Paris auctions, 1764 - 1780",
    x = "Width (inches)",
    y = "Height (inches)"
  )
```

Other smoothing methods: gam

Plot Code

```
ggplot(data = pp, aes(x = Width_in, y = Height_in)) +
  geom_point() +
  geom_smooth(method = "gam",
              se = FALSE, color = "#8E2C90") +
  labs(
    title = "Height vs. width of paintings",
    subtitle = "Paris auctions, 1764 - 1780",
    x = "Width (inches)",
    y = "Height (inches)"
  )
```

Other smoothing methods: loess

Plot Code

```
ggplot(data = pp, aes(x = Width_in, y = Height_in)) +
  geom_point() +
  geom_smooth(method = "loess",
              se = FALSE, color = "#8E2C90") +
  labs(
    title = "Height vs. width of paintings",
    subtitle = "Paris auctions, 1764 - 1780",
    x = "Width (inches)",
    y = "Height (inches)"
  )
```

Vocabulary

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

Residuals

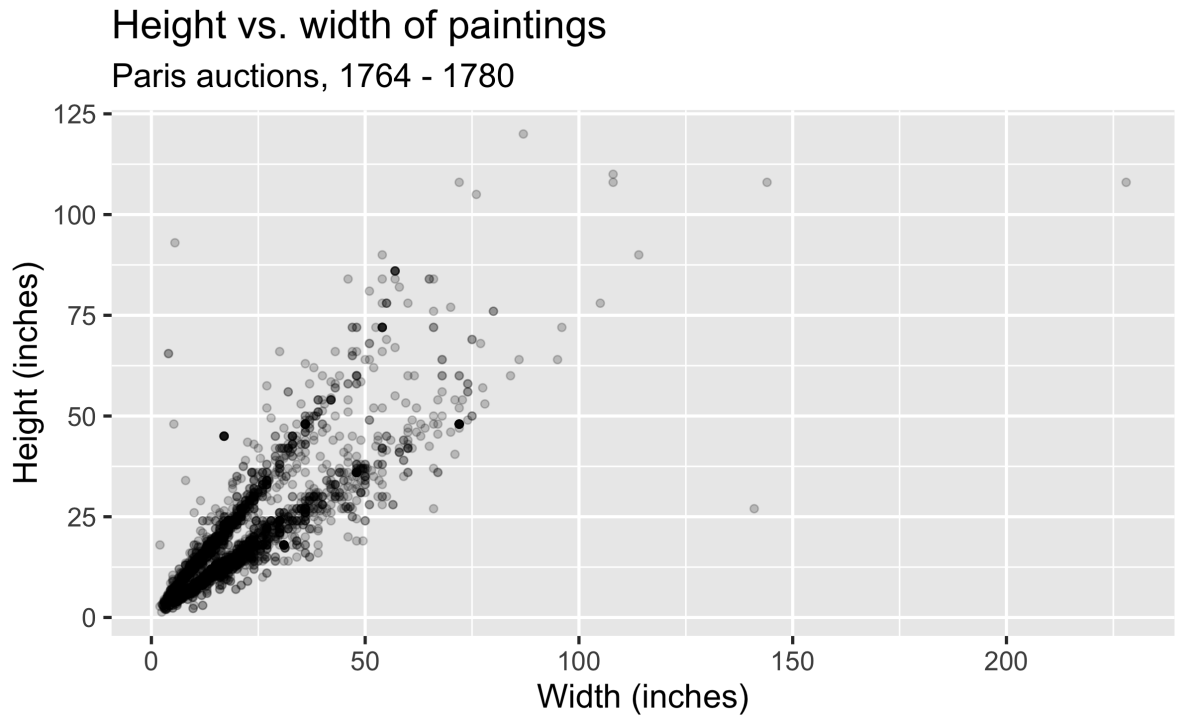
Plot Code

```
ht_wt_fit <- linear_reg() %>%
  set_engine("lm") %>%
  fit(Height_in ~ Width_in, data = pp)

ht_wt_fit_tidy <- tidy(ht_wt_fit$fit)
ht_wt_fit_aug  <- augment(ht_wt_fit$fit) %>%
  mutate(res_cat = ifelse(.resid > 0, TRUE, FALSE))

ggplot(data = ht_wt_fit_aug) +
  geom_point(aes(x = Width_in, y = Height_in, color = res_cat)) +
  geom_line(aes(x = Width_in, y = .fitted), size = 0.75, color = "#8E2C90") +
  labs(
    title = "Height vs. width of paintings",
    subtitle = "Paris auctions, 1764 - 1780",
    x = "Width (inches)",
    y = "Height (inches)"
  ) +
  guides(color = FALSE) +
  scale_color_manual(values = c("#260b27", "#e6b0e7")) +
  geom_text(aes(x = 0, y = 150), label = "Positive residual", color = "#e6b0e7", hjust = 0, size = 8) +
  geom_text(aes(x = 150, y = 25), label = "Negative residual", color = "#260b27", hjust = 0, size = 8)
```

The plot below displays the relationship between height and width of paintings. The only difference from the previous plots is that it uses a smaller alpha value, making the points somewhat transparent. 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 they are long.
- Portrait painting is a genre in painting, where the intent is to depict a human subject.[2]
  - Portrait paintings tend to be longer than they are wide.

[1] Source: Wikipedia, Landscape painting

[2] Source: Wikipedia, Portait painting

Multiple explanatory variables

Plot 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 = "Height vs. width of paintings, by landscape features",
    subtitle = "Paris auctions, 1764 - 1780",
    x = "Width (inches)",
    y = "Height (inches)",
    color = "landscape"
  ) +
  scale_color_manual(values = c("#E48957", "#071381"))
```

Extending regression lines

Plot 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 = "Height vs. width of paintings, by landscape features",
    subtitle = "Paris auctions, 1764 - 1780",
    x = "Width (inches)",
    y = "Height (inches)",
    color = "landscape"
  ) +
  scale_color_manual(values = c("#E48957", "#071381"))
```

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?

- Explanation: Characterize the relationship between y and x via *slopes* for numerical explanatory variables or *differences* for categorical explanatory variables
- Prediction: Plug in x, get the predicted y