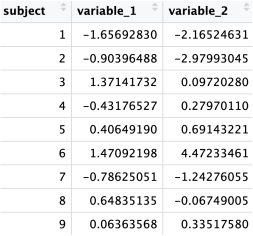
Welcome to STA 101!

Statistics is a confrontation with uncertainty.

Statistics confronts uncertainty by quantifying it.

Data analysis

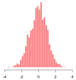
Transforming messy, incomplete, imperfect data into knowledge:

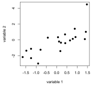


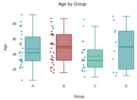
subject	variable_1	variable_2
1	-1.65692830	-2.16524631
2	-0.90396488	-2.97993045
3	1.37141732	0.09720280
4	-0.43176527	0.27970110
5	0.40649190	0.69143221
6	1.47092198	4.47233461
7	-0.78625051	-1.24276055
8	0.64835135	-0.06749005
9	0.06363568	0.33517580

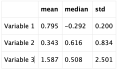
`ggplot` ⟹

`|>` ⟹









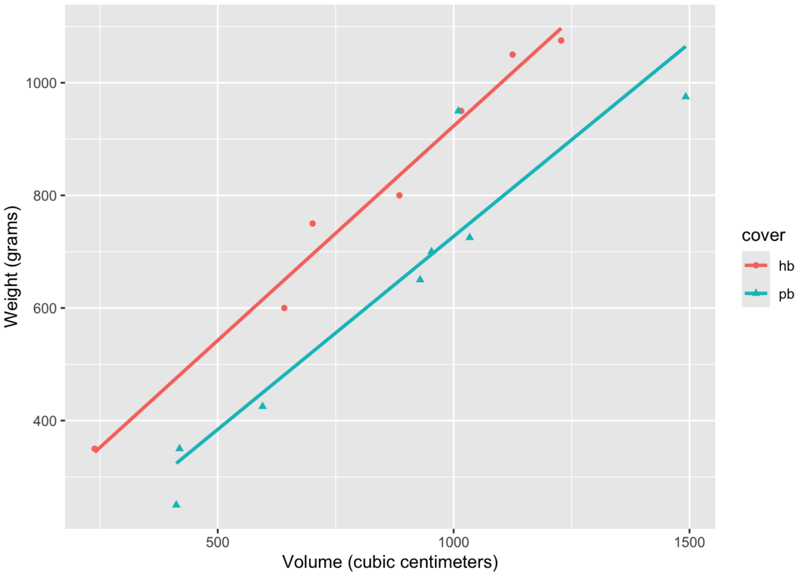
	mean	median	std
Variable 1	0.795	−0.292	0.200
Variable 2	0.343	0.616	0.834
Variable 3	1.587	0.508	2.501

Statistical inference

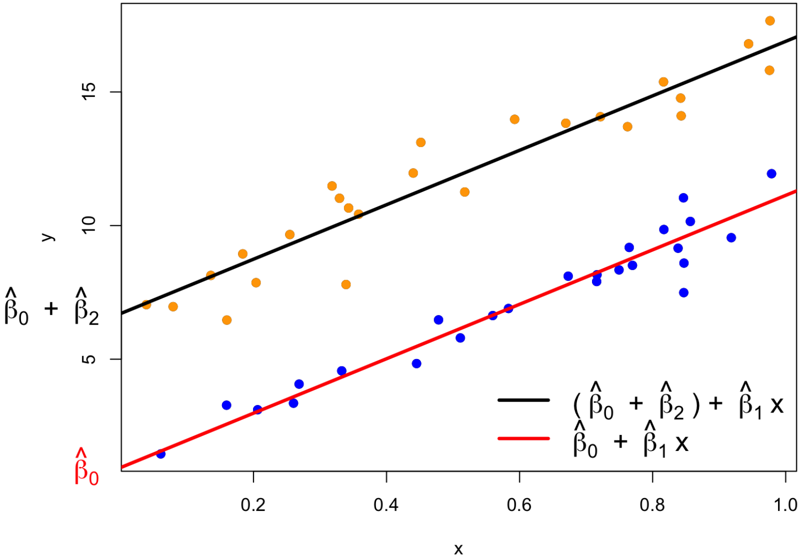
Quantifying our uncertainty about that knowledge:

- **Question**: What's the number?
- **Answer**: best-guess ± margin-of-error

The book example from the primer



One model, but two lines?



One model, but two lines?

We fit a model with a numerical predictor (x_1) and a categorical predictor (x_2) with two level:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2.$$

The categorical predictor works like this:

$$x_2 = \begin{cases} 0 & \text{if book is hardback} \\ 1 & \text{if book is paperback.} \end{cases}$$

This essentially nests two models:

$$\hat{y} = \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 x_1 & \text{if book is hardback} \\ \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 & \text{if book is paperback.} \end{cases}$$

The intercept shifts.

(JZ: In the book example, $\hat{\beta}_2$ is negative, so it shifts the line down.)

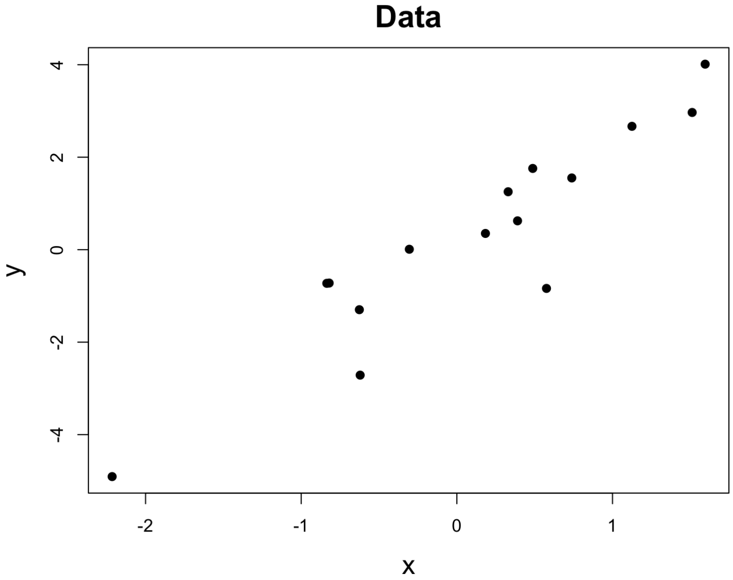
What are models good for?

"All models are wrong but some are useful."

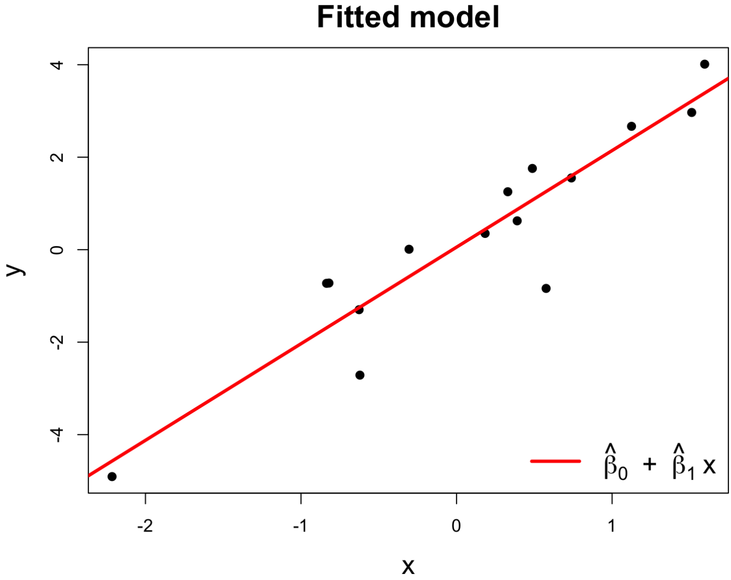
- We would love to use models to draw causal conclusions from messy, observational data, but this is very very difficult;

- A slightly easier task that some models are very good at is *prediction*...

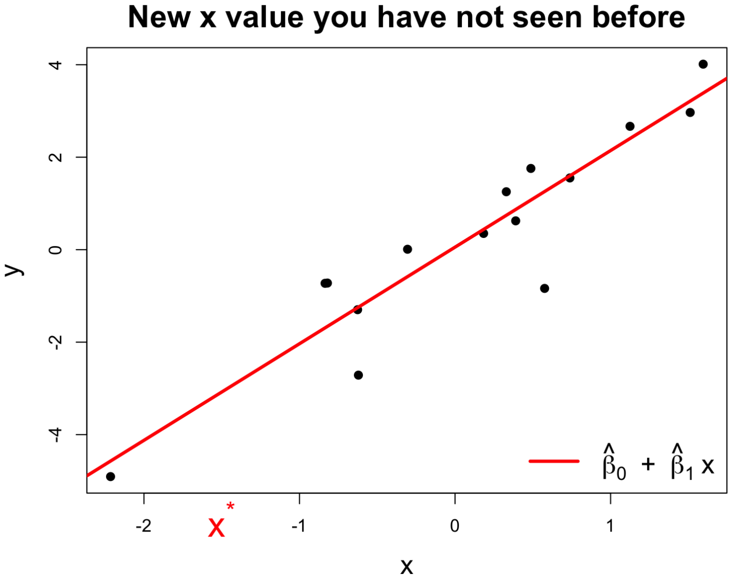
Prediction



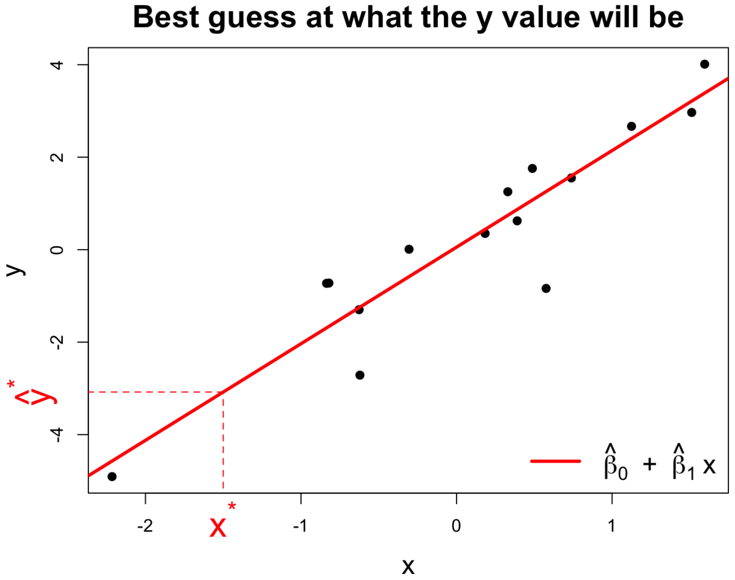
Prediction



Prediction


New x value you have not seen before
y
4
2
0
-2
-4
x
-2
x^*
-1
0
1
$\hat{\beta}_0 + \hat{\beta}_1 x$

Prediction



Best guess at what the y value will be
y
4
2
0
-2
-4
$\hat{y}^*$
x
-2
x^*
-1
0
1
$\hat{\beta}_0 + \hat{\beta}_1 x$

Prediction, summary

Collect data:

$$(x_1,\ y_1), (x_2,\ y_2), (x_3,\ y_3), ..., (x_n,\ y_n)$$

Fit a model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

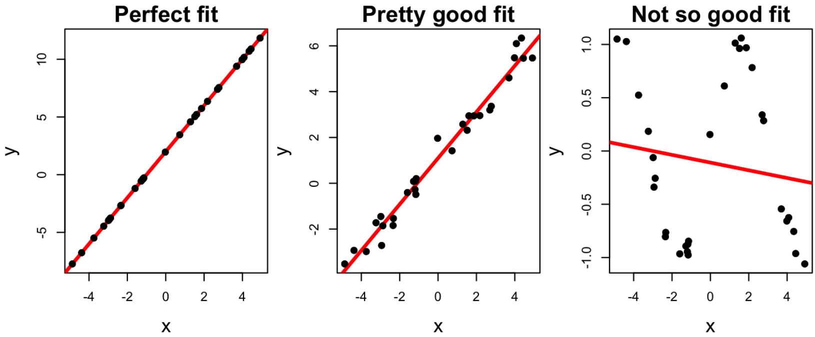
Prediction: Someone hands you an $x^\star$ you've never seen before. What's your best guess at what y is going to be?

$$\hat{y}^\star = \hat{\beta}_0 + \hat{\beta}_1 x^\star.$$

Historical Data + Model = Predictions you can use to aid decision making in an uncertain world.

(JZ: In the book example, you use the model to guesstimate the weight of a book you are about to manufacture. Maybe this allows you to preallocate the budget for the shipping costs. You can't wait until the book is manufactured to decide on your budget. It must be done now, and the data and model help you predict and resolve some of your uncertainty about future costs.)

How well does a linear model fit the data?



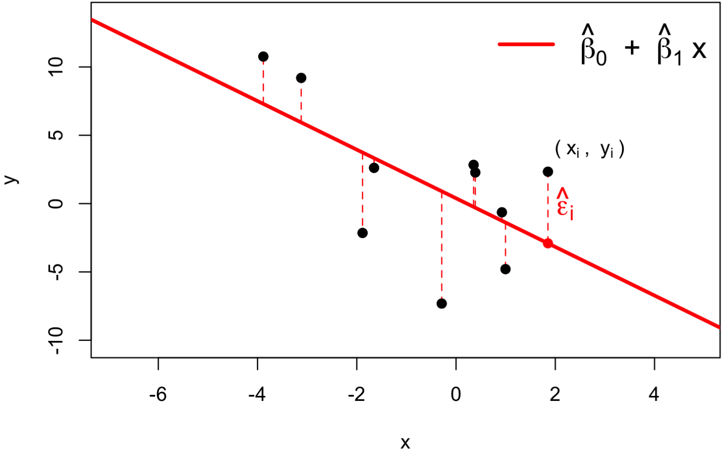
Question: can we *quantify* this?

(JZ: Visually, fit would seem to have *something* to do with how far the data points are from the fitted line. In other words, how big the residuals are. So we need to talk about residuals...)

Recall the residuals

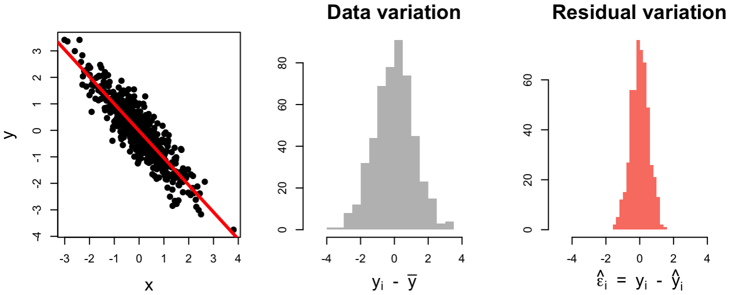
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad \text{(fitted model)}$$

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \qquad \text{(residuals)}$$



How variable are the residuals compared to the data?

Idea: Fit a model, and compare histogram of $\{y_1, y_2, \ldots, y_n\}$ to histogram of $\{\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n\}$:

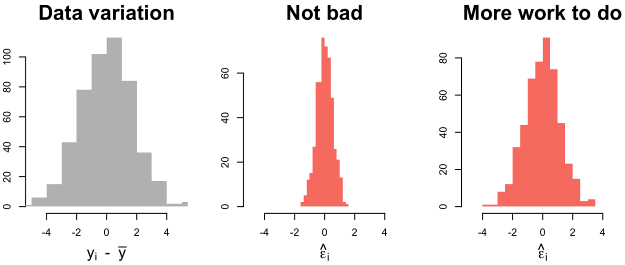


- Variation in y_i is what we seek to "explain" with a model;
- Variation in $\hat{\varepsilon}_i$ is the leftover that our model does not explain;
- If there's not a lot of leftover, we did pretty well.

How variable are the residuals compared to the data?

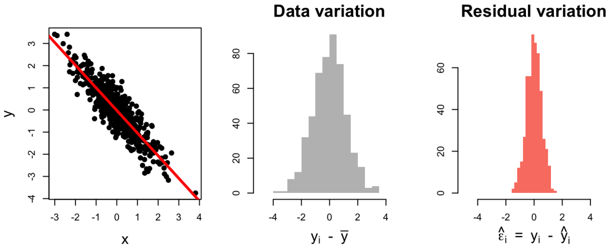
Idea: Fit a model, and compare histogram of $\{y_1, y_2, ..., y_n\}$ to histogram of $\{\hat{\varepsilon}_1, \hat{\varepsilon}_2, ..., \hat{\varepsilon}_n\}$.

- Variation in y_i is what we seek to "explain" with a model;
- Variation in $\hat{\varepsilon}_i$ is the leftover that our model does not explain;
- If there's not a lot of leftover, we did pretty well:



(JZ: If the residual distribution is a point mass at zero with no variation, we explained everything. Winner! If the residual distribution looks identical to the data distribution, we explained nothing. Drat!)

How do we quantify this comparison?



Summarize variation with a measure of *spread* and compare:

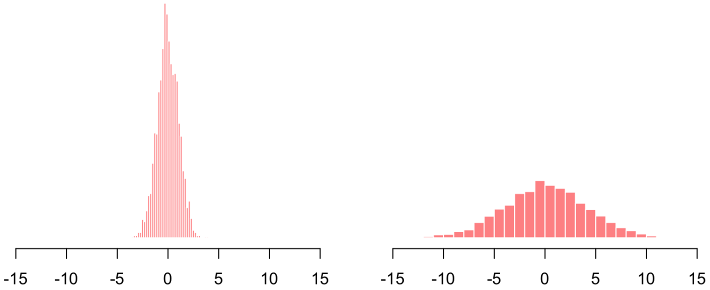
$$
\begin{aligned}
\text{fit quality} &= \text{proportion of variation explained} \\
&= \frac{\text{explained variation}}{\text{total data variation}} \\
&= \frac{\text{total data variation - unexplained variation}}{\text{total data variation}} \\
&= \frac{\text{spread of } y_i - \text{spread of } \hat{\varepsilon}_i}{\text{spread of } y_i}.
\end{aligned}
$$

How do we measure spread?

With the **variance**:

the average squared distance from the mean.

“how far are the data, typically, from their center?”



Left: data are typically close to the center (low variance)

Right: data are typically farther from the center (higher variance).

How do we measure spread?

With the **variance**:

the average squared distance from the mean.

“how far are the data, typically, from their center?”

Recall that the mean is the same thing as the average:

$$\bar{y} = \frac{y_1 + y_2 + y_3 + \ldots + y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i .$$

So in formulas...

$$\begin{aligned}
\operatorname{var}(y_i) &= \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \operatorname{sd}(y_i)^2. \quad \text{(“variance is standard deviation squared”)}
\end{aligned}$$

And similarly for $\operatorname{var}(\hat{\varepsilon}_i)$

Back to measuring fit

Before:

$$\begin{aligned}\text{fit quality} &= \text{proportion of variation explained by the model} \\ &= \frac{\text{spread of } y_i - \text{spread of } \hat{\varepsilon}_i}{\text{spread of } y_i}.\end{aligned}$$

Now:

$$\begin{aligned}R^2 &= \frac{\text{var}(y_i) - \text{var}(\hat{\varepsilon}_i)}{\text{var}(y_i)} \\ &= 1 - \frac{\text{var}(\hat{\varepsilon}_i)}{\text{var}(y_i)}.\end{aligned}$$

R^2 is called the **coefficient of determination**.

Facts about R^2

- $\text{var}(y_i)$ is always bigger than $\text{var}(\hat{\varepsilon}_i)$, so R^2 is a number between zero and one;

- If $\text{var}(\hat{\varepsilon}_i) = 0$, then the model explained everything. The fit is perfect (poifect!), and $R^2 = 1$;

- If $\text{var}(\hat{\varepsilon}_i) = \text{var}(y_i)$, then the model explained absolutely nothing and $R^2 = 0$;

- Most of the time we are somewhere in between, and we can use R^2 to quantify the quality of a model's fit and *rank* competing models.

Adjusted R^2

Possible use of R^2:

- decide which covariates to include in a big multiple regression. The set of covariates that delivers the highest R^2 is the winner;

Problem:

- R^2 has a nasty mathematical property that it *always goes up* every time you add *any* covariate to the model, even if that covariate is silly and useless;

Goal:

- We want a measure of fit that will not give all variables a participation trophy just for showing up, but actually rewards honest-to-goodness improvements in fit;

Solution:

- Adjusted R^2.

Variable selection: backward elimination

Start with the *full model* (the model that includes all potential predictor variables). Variables are eliminated one-at-a-time from the model until we cannot improve the model any further.

Procedure:

1. Start with a model that has all predictors we consider and compute the adjusted R^2.
2. Next fit every possible model with 1 fewer predictor.
3. Compare adjusted R^2s to select the best model (highest adjusted R^2) with 1 fewer predictor.
4. Repeat steps 2 and 3 until adjusted R^2 no longer increases.

Variable selection: forward stepwise

Forward stepwise regression is the reverse of the backward elimination technique. Instead, of eliminating variables one-at-a-time, we add variables one-at-a-time until we cannot find any variables that improve the model any further.

Procedure:

1. Start with a model that has no predictors.
2. Next fit every possible model with 1 additional predictor and calculate adjusted R^2 of each model.
3. Compare adjusted R^2 values to select the best model (highest adjusted R^2) with 1 additional predictor.
4. Repeat steps 2 and 3 until adjusted R^2 no longer increases.