Chapter 5.2 - Least Squares

"On two occasions I have been asked [by members of Parliament], 'Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?' I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question." - Charles Babbage

The best part about being a mathematician is listening to everyone you know complain about how painful their education in the slope-intercept formula was, knowing that it's the simplest equation math has to offer.

The best part about being a statistician is listening to everyone you know complain about how useless their education in the slope-intercept formula was, knowing that it's the most useful equation in history.

Modern statisticians sit on the shoulders of dead European men who like looking at stars. In 1805, Adrien-Marie Legendre described the method of least squares as a technique estimating the orbit of comets. Another astronomer, Carl Friedrich Gauss, would go on to define the method in the framework of statistics. Little did Gauss know, (although he probably suspected since statistcians are notoriously self-obsessed), he had produced the greatest contribution to science, at the time of writing this, any human being has ever made.

A strange thing to say considering this was his victory lap after defining *the normal distribution*. Even stranger when you see the formula I'm referring to:

$$y = \beta_0 + \beta_1 x + \epsilon$$

It should look *eerily* familiar if we redefine $\beta_0 = b$ and $\beta_1 = m$:

$$y = b + mx + \epsilon$$

Drop the ϵ term then rearrange the equation and you'll get:

$$y = mx + b$$

The slope-intercept formula. How is *this* the most important model in history? Least squares regression is a small expansion to this formula that includes residual error, which can be transformed into the **linear model** with the addition of a single assumption. With the linear model we put a man on the moon, predicted natural disasters, and developed machine learning (popularized by "artificial intelligence" models).

In this section we'll discuss how to use the simplest form of this model, the general theory behind it, and the assumptions we have to meet (or at the very least accept) in order to capitalize on its power.

Estimating β

When we say the "method of least squares" we're typically referring to **least squares regression** (LSR); sometimes called **ordinary least squares** (OLS). The formula for LSR is given as:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Where y_i is the response variable, β_0 is the intercept term, β_1 is the slope, x_i is the predictor variable, and ϵ_i is residual error. The table below describes the rough interpretation of each variable.

Variable	Name	Interpretation
y	Response	The variable being predicted
x	Predictor	The variable predicting the response
β_0	Intercept	The baseline level of the response
β_1	Slope	The effect of the predictor on the response
ϵ	Residual error	The difference between the regression line and the data

LSR is a deterministic, implicit, phenomenological model. So we know that any inputs we give it will produce the same result every time, the way that we make use of it is by estimating the unknown parameters, and it's helping us understand a process using data rather than using a process to understand data.

There are two common techniques for "fitting", or estimating the parameters of, LSR: matrix and scalar technique. We'll cover the scalar technique and leave the explanation of matrix technique for another time.

We start by estimating β_1:

$$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2}$$

Whenever we estimate a variable we denote it with a "hat" symbol (ˆ) to differentiate between the true parameter and the estimated one. With $\hat{\beta}_1$ we can solve for β_0:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

With both parameters estimated we can plug in the observed values for x to the *explicit* regression equation and estimate y:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Then with $\hat{y}$ we can estimate ϵ:

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

In order to use least squares we need at least 3 data points, since 1 point is fully described by its coordinates and 2 would just need a steady hand to draw a line through.

We'll test these formulas on the trivial data set below:

x	y
1	0.16
2	2.82
3	2.24

Let's start by calculating the two sample means:

$$\bar{x} = \frac{1+2+3}{3} = \frac{6}{3} = 2$$

$$\bar{y} = \frac{0.16+2.82+2.24}{3} = \frac{5.22}{3} = 1.74$$

A good method for attacking the next steps is to build a table for all of the intermediary calculations then calculate the remaining summations:

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
-1	-1.58	1.58	1
0	1.08	0	0
1	0.5	0.5	1

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = 1.58 + 0 + 0.5 = 2.08$$

$$\sum_i (x_i - \bar{x})^2 = 1 + 0 + 1 = 2$$

And we can finish with the classic "plug and chug".

$$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2} = \frac{2.08}{2} = 1.04$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 1.74 - 1.04 \times 2 = -0.34$$

These values for $\hat{\beta}_0$ and $\hat{\beta}_1$ form the regression equation:

$$\hat{y} = -0.34 + 1.04 x_i$$

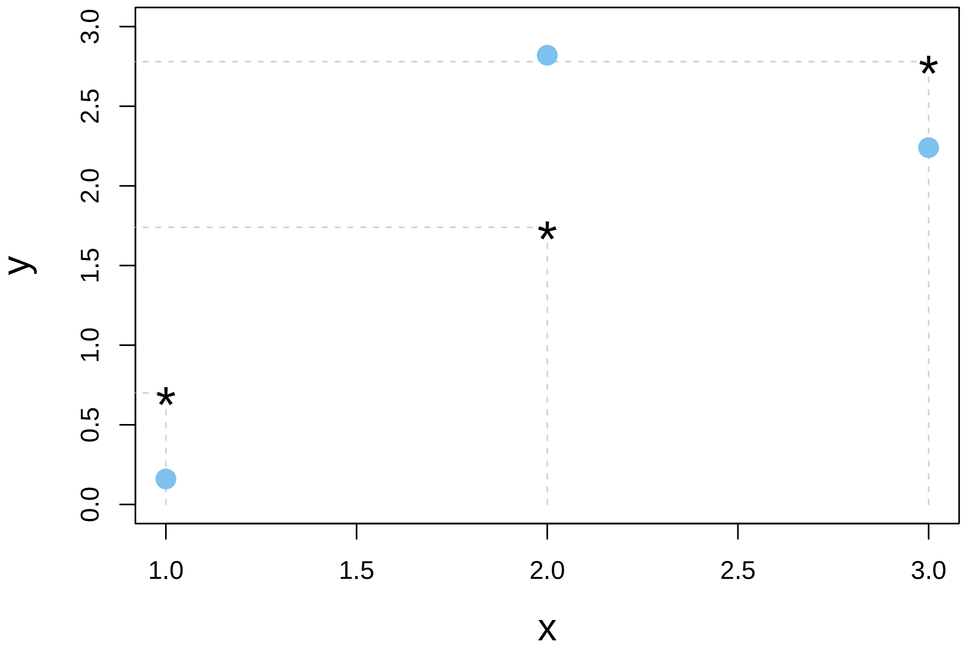
We can plug in all of the values for x to calculate $\hat{y}$:

$$\hat{y}_1 = -0.34 + 1.04 \times 1 = 0.7$$

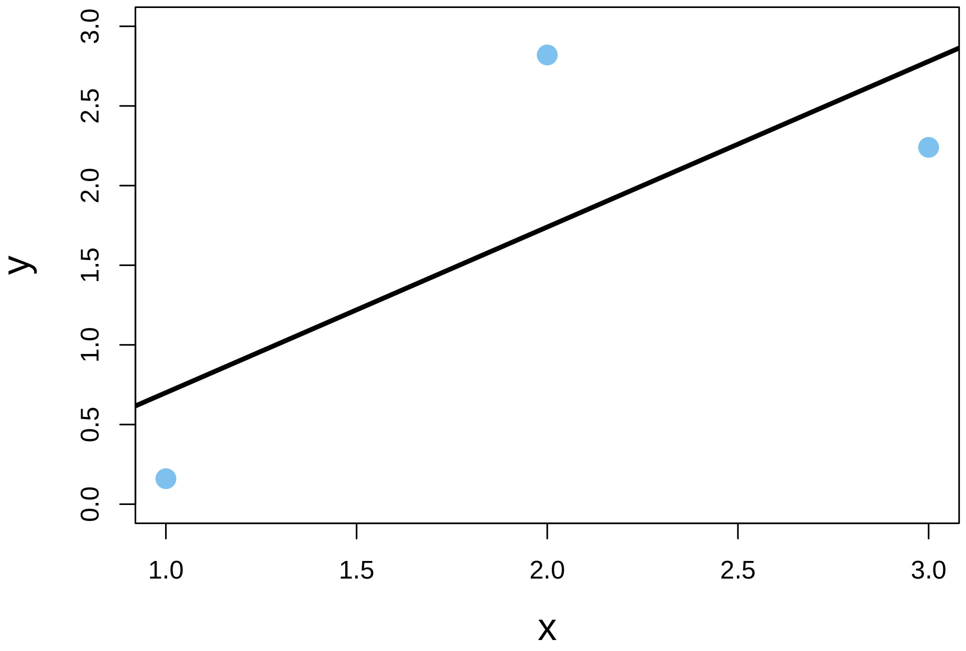
$$\hat{y}_2 = -0.34 + 1.04 \times 2 = 1.74$$

$$\hat{y}_3 = -0.34 + 1.04 \times 3 = 2.78$$

Then we're done! We can plot these predicted points alongside the original data:



We can see how these can be connected to form a line showing the relationship between x and y.



While we should never lean on computational programs to check our work, it should be quite comforting to learn that these calculations match those we can produce in R using the `lm()` function (an analog to LSR when we only look at the coefficients):

```
coef(lm(Y~X))
```

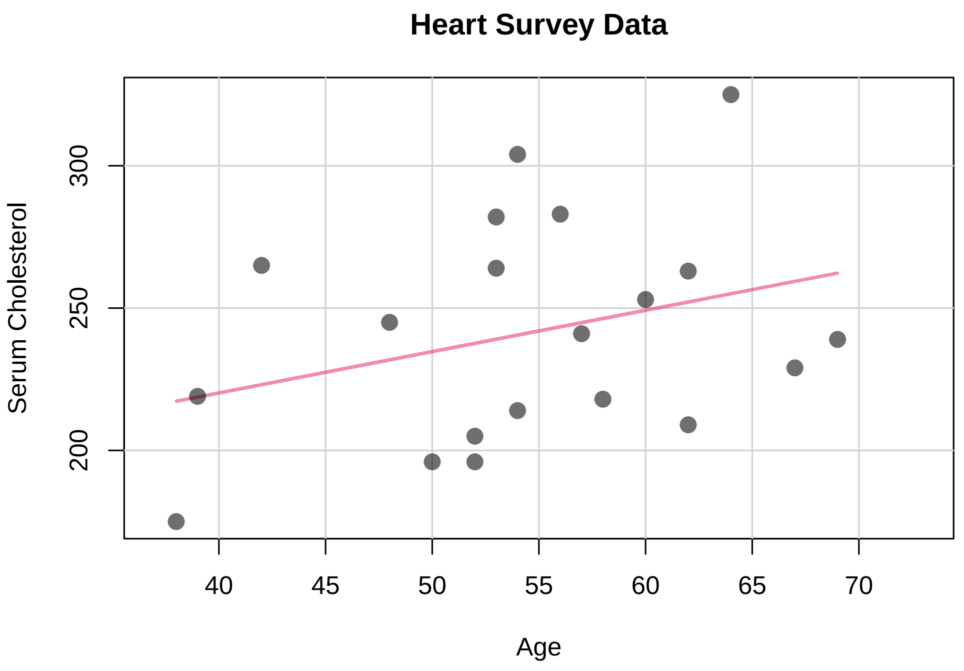
```
## (Intercept)           X
##       -0.34        1.04
```

In practice we prefer to use programs like R for fitting LSR to data since we rarely have an amount of data that's reasonable to work with by hand. It's still good to demystify the underlying functions that give us the regression line so that we can comfortably work with the output (and recognize when LSR isn't going to cut it).

Loss Functions

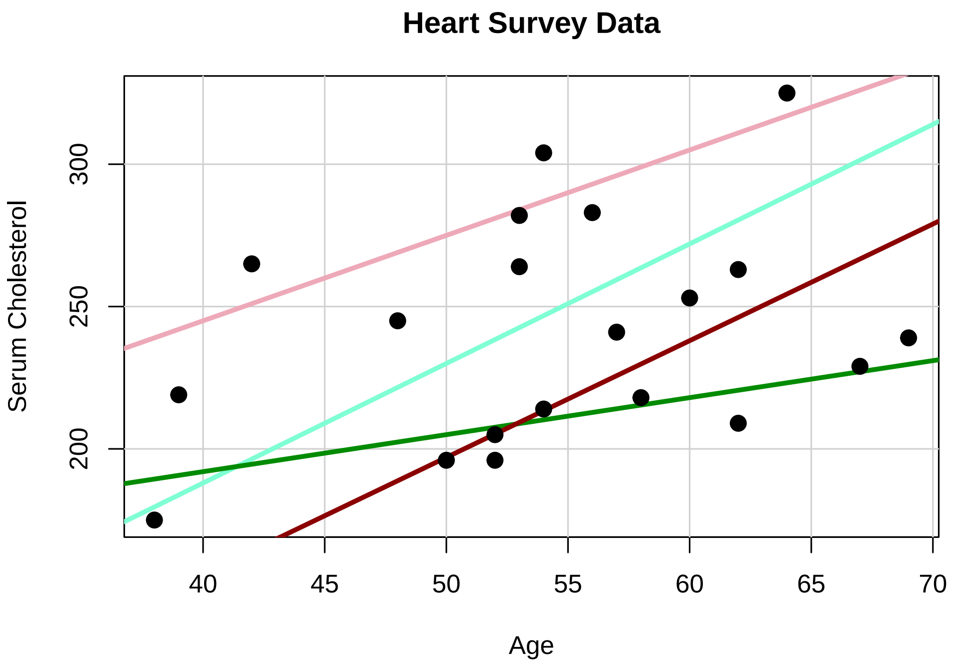
I didn't avoid estimating $\hat{\epsilon}_i$ in the last section because I felt it was unimportant or because I forgot (although that's a frequent problem in my life). We need to discuss what "residual error" truly means before we can proceed with quantifying it.

Let's start with this: when we fit a least squares regression line we're making a mathematical simplification of reality. A simplification that contains inaccuracies that statisticians refer to as "error".



But how do we decide which inaccuracies are acceptable? What do we do with the calculated errors? What makes this *the* least squares regression line and why wouldn't a different line work?

First and foremost, there are other *viable* lines:



But none of these are the *least squares* regression line (also, none of those lines are the result of models but that's beside the point). When we use the *method* of least squares we're fitting the model via a **loss function**.

> You've been set up on a blind dinner date by some friends. You wait around outside the restaurant until they eventually arrive— 10 minutes late. You head inside and get a table. The restaurant's nice, they're cute, the conversations are fun, they're nice to the wait staff.
>
> The waiter comes up to take your orders but before you can get a word out your date orders for both of you. They tell you they're a regular and know all the best things to get. They just wanted to give you the best experience possible. The food, despite being nothing you would have thought to order, turns out to be amazing.
>
> You finish your food, the bill comes out. You decide to split the cost, even though it's much more than you would have paid if you'd ordered your original choice. You see your date write a 0 in the tip line. As you're walking out they comment that this date went much better their first date with their ex.

At some point in this date you made a decision on whether or not you were texting this person back. Unbeknownst to you, this was probably the product of how harshly you judged the red flags that came up during the date. The level of disgust you felt at your date's different actions is analogous to how a loss function penalizes errors in a model.

Loss functions are a sort of mathematical "rule" that we can measure our errors with to help us determine what's "acceptable".

Least squares uses *squared error* loss:

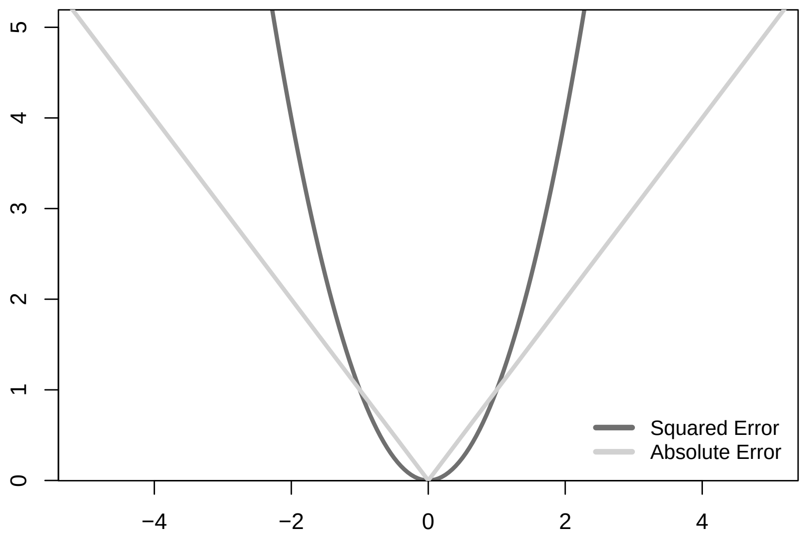
$$L(x, a) = (x - a)^2$$

Squared loss arises from the same general idea as sample variance; if we didn't square the difference between the "true" values and the predictions we would have to deal with negatives. Another option to deal with negative values would be to use absolute values, this would be called *absolute error* loss:

$$L(x, a) = |x - a|$$

The same logic exists for both loss functions as with variance and absolute differences. Squared loss can be fit by hand while absolute loss isn't. But as we use computational programs to work with most models there has to be more rationale behind choosing loss functions than analytical simplicity.

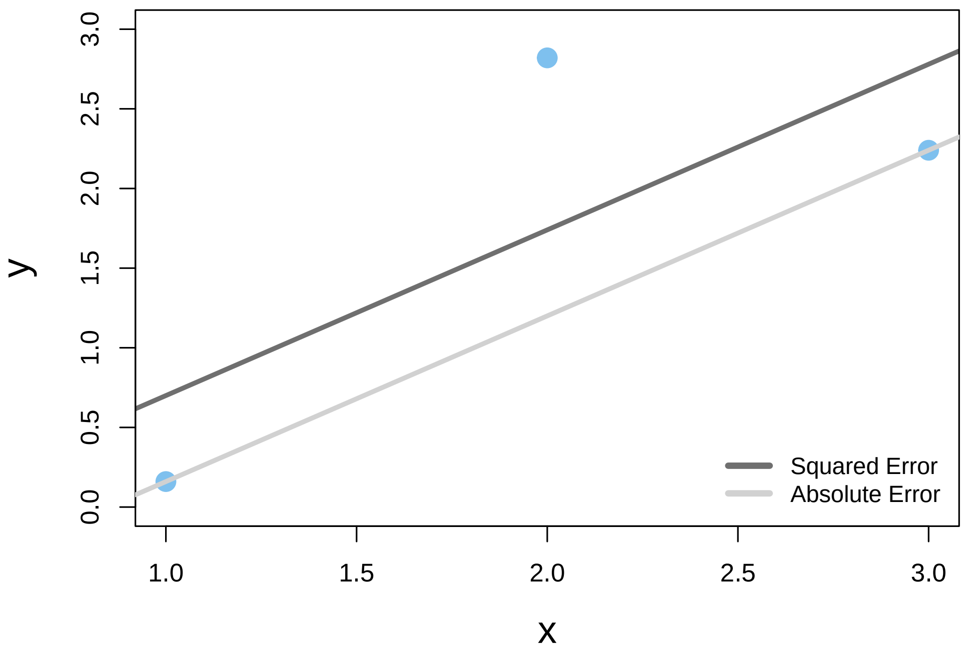
Squared loss penalizes error *quadratically*, larger errors are giving higher value than small one. Absolute loss penalizes error *linearly*, all error contributes equally.



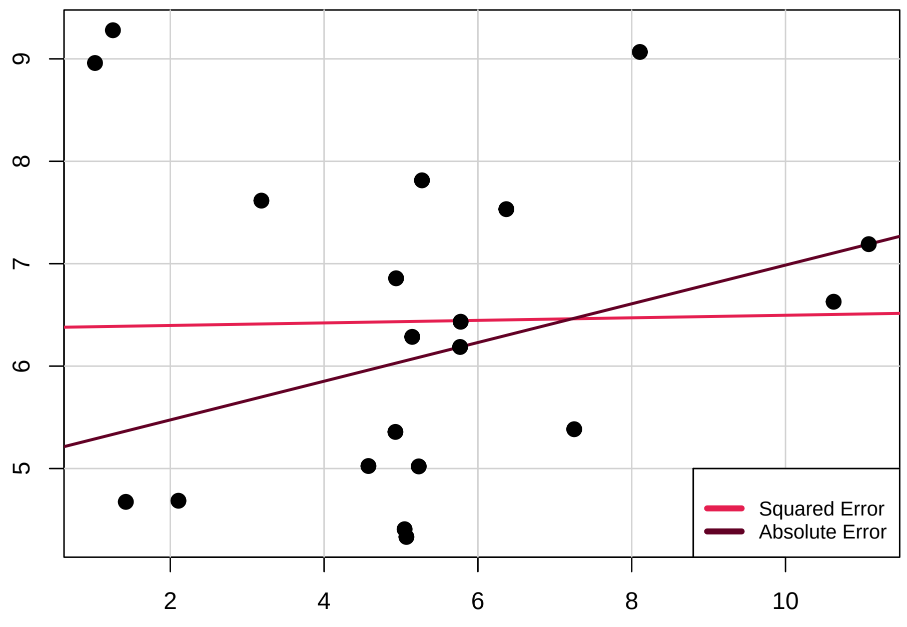
We can fit a regression model using absolute error loss, commonly called *least absolute deviations.* Using the trivial, $n = 3$ data set from before as an example:

x	y
1	0.16
2	2.82
3	2.24

	β_0	β_1
Squared Error	-0.34	1.04
Absolute Error	-0.88	1.04



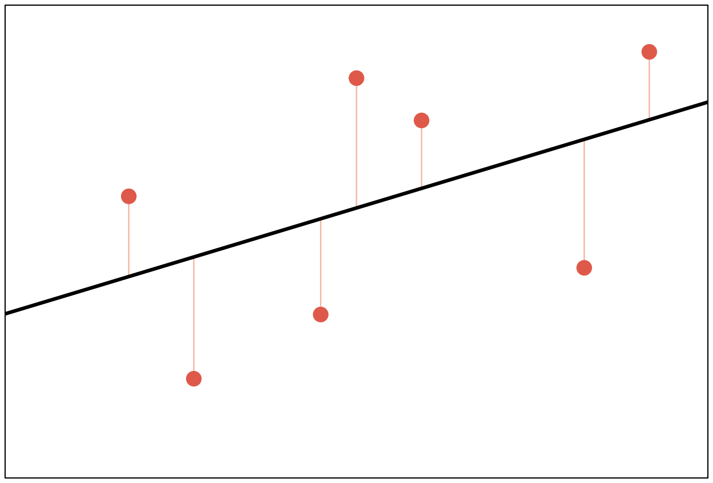
As a result of linear penalization of error, absolute loss will fit a line closer to the *densest* region of data rather than trying to minimize the distance between all of the data. On occasion this will provide completely different results:



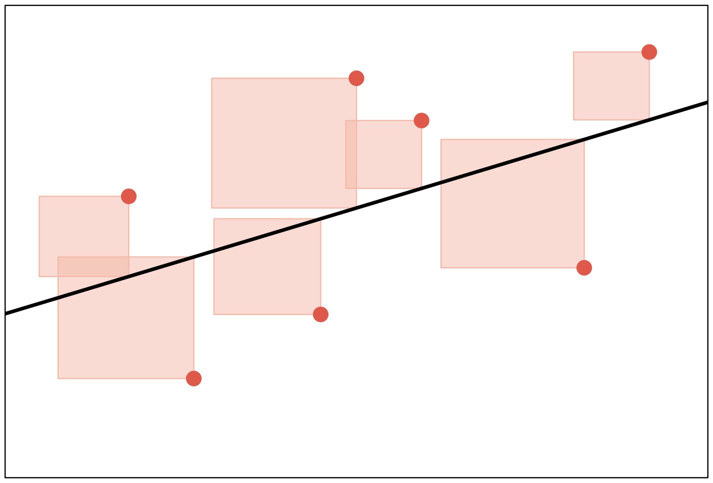
But more often than not the shift will be minor.

This idea of squared distance may have felt strange when we discussed it in the context of variance but for least squares we can develop a much more intuitive explanation.

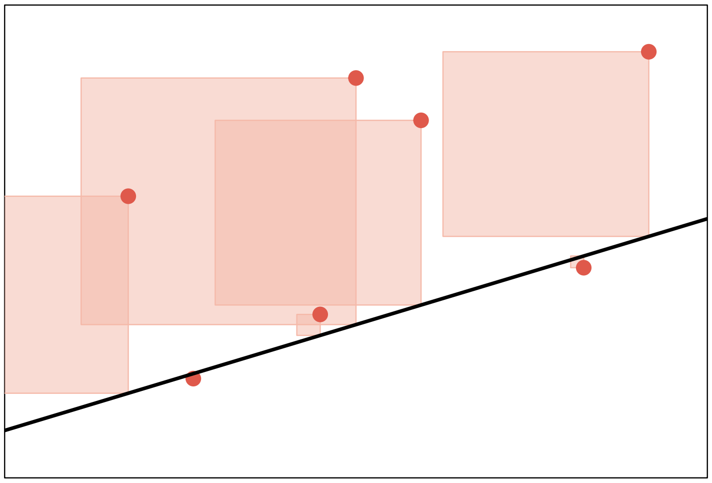
The distances between the true values of the data and the regression line are the residuals.



If we tried to minimize the total length of these lines when stacked on top of each other we would be using absolute loss. With least squares we go one step further and stretch those lines into squares.



The objective of squared loss is to *minimize* the total sum of those squares. Even though some lines would allow for certain points to produce squares with near zero area, far away points increase the sum of those squares quadratically.



This is why least squares is so useful because rather than focus on getting as many predictions "correct" as possible, as is sometimes the goal with least absolute deviations, the natural objective of least squares is to get a minimal number of "egregiously wrong" predictions.

Assumptions

Let's return to our favorite $n = 3$ data set with the addition of the predicted values of the response, $\hat{y}$.

x	y	$\hat{y}$
1	0.16	0.7
2	2.82	1.74
3	2.24	2.78

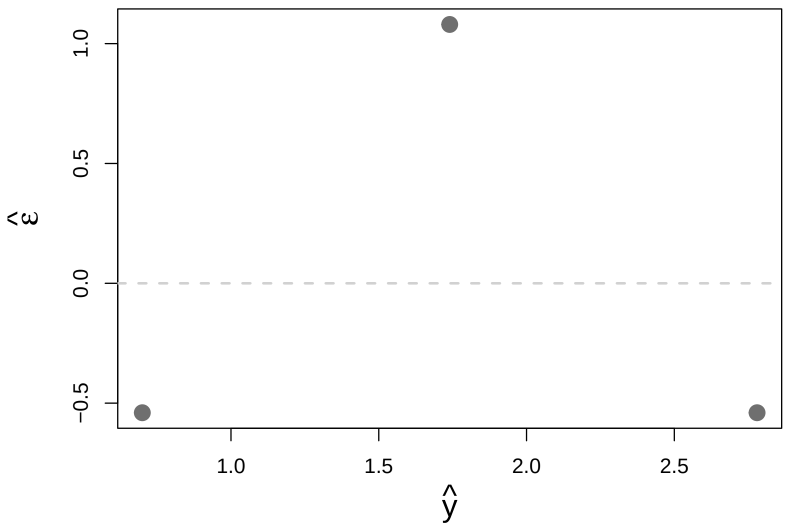
To estimate the residuals, $\hat{\epsilon}_i$, we subtract the observed values of y from the predicted values.

x	y	$\hat{y}$	$y - \hat{y}$
1	0.16	0.7	-0.54
2	2.82	1.74	1.08
3	2.24	2.78	-0.54

These residuals have a few fun properties that define the inherent *assumptions* of least squares. The first is that the the residuals are **independent** from one another, which we can verify by showing that they have a mean of 0.

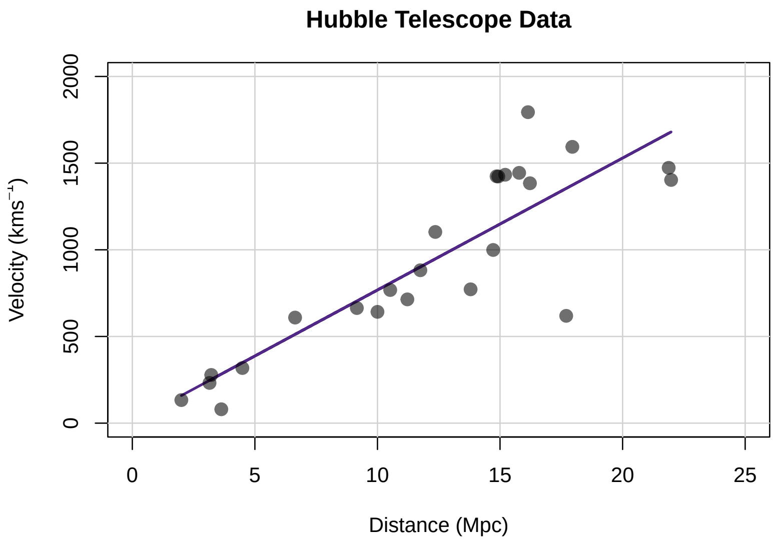
$$\bar{\epsilon} = \frac{\hat{\epsilon}_1 + \hat{\epsilon}_2 + \hat{\epsilon}_3}{3} = \frac{-0.54 + 1.08 - 0.54}{3} = \frac{0}{3} = 0$$

The next is that they have constant variance or **homoscedasticity**. There are more formal methods of verifying this but plotting the residuals, $\hat{\epsilon}$, against the fitted values, $\hat{y}$, is more than enough in most scenarios.



The last is the assumption of **linearity**, a concept we've already discussed. The data doesn't need to cluster on a perfect line (or even very closes around a line) but we shouldn't try to use least squares on an obviously non-linear data set.

It's good practice to "check" the assumptions of our models at some point in the analysis (ideally before the presentation of results). We'll walk through this process with the hubble telescope data.

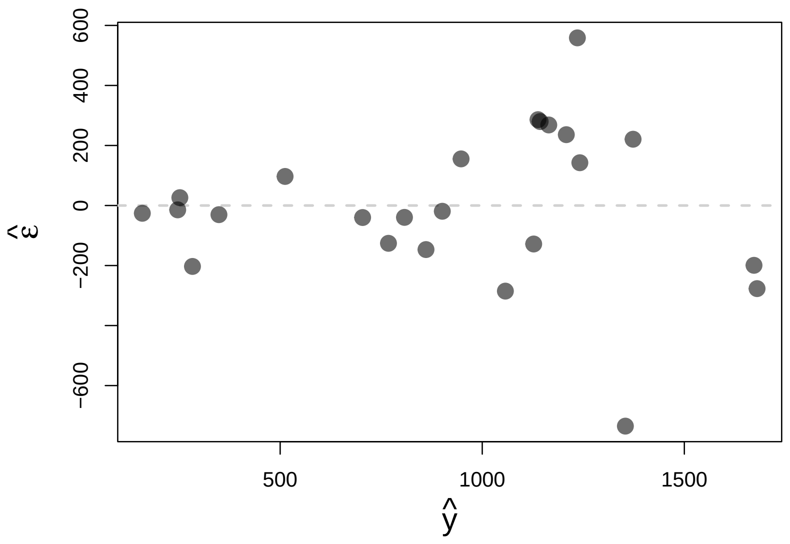


Looking at the plot alone should be enough to form an argument that the assumption of linearity is met. Checking the assumption of independence in the residuals is possible by hand but we'll use R given how much data we're dealing with:

```r
# calculate the mean and round to 10 decimal places
round(mean(resids),10)
```

```
## [1] 0
```

Then we can plot the residuals to check for homoscedasticity:



This plot should make us a little uncomfortable. It seems the variance in the residuals increase as $\hat{y}$ increases, which coincides with the observations that Hubble is recording coming from further away. We have reason to suggest that the assumption of homoscedasticity isn't *approximately met*, but what do we do now?

When our assumptions haven't been approximately met there are two standard options for next steps:

1. Adjust the model so as to adjust the assumptions.
2. Continue using the same model but publicly disclose the inaccurate assumptions.

Statistics, as a science, is unconcerned with whether or not a method/model/answer is wrong so long as it's made clear (and ideally quantified). If we don't want to proceed with using the model because the risk of a wrong answer is too great then we can seek out a statistician and have them help us develop a new one. If what we're looking to do with our model doesn't depend on supreme predictive accuracy or perfectly capturing the data generating process then we should use the model *we want to use*.