Least Squares Regression

October 30, 2019

We want a line with small residuals, so it might make sense to try to minimize

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

...but this will give us very large negative residuals!

Finding the Best Line

As with the standard deviation, we will use squares to shift the focus to magnitude:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Finding the Best Line

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

The values of b that minimize this will make up our regression line.

This is called the **Least Squares Criterion**.

Finding the Best Line

To fit a **least squares regression**, we require

- Linearity. The data should show a linear trend.
- Nearly normal residuals. The residuals should be well-approximated by a normal distribution.
- Constant variability. As we move along x, the variability around the regression line should stay constant.
- Independent observations. This will apply to random samples.

Finding the Least Squares Line

We want to estimate β_0 and β_1 in the equation

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

by minimizing $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Finding the Least Squares Line

This turns out to be remarkably straightforward! The slope can be estimated as

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

and the intercept by

$$b_0 = \bar{y} - b_1 \bar{x}$$

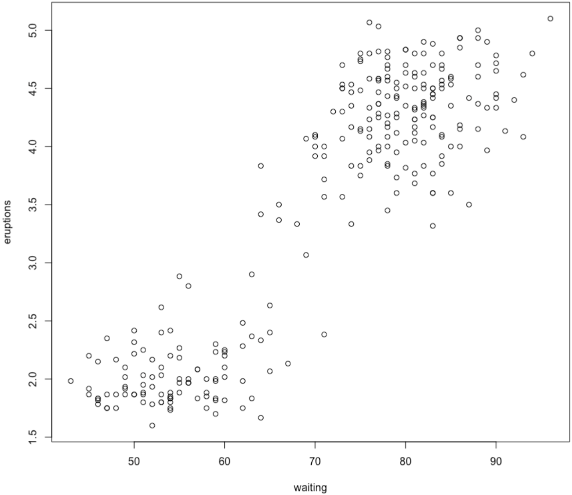
Example

The `faithful` dataset in `R` has two measurements taken for the Old Faithful Geyser in Yellowstone National Park:

- `eruptions`: the length of each eruption
- `waiting`: the time between eruptions

Each is measured in minutes.

Example



We want to see if we can use the wait time to *predict* eruption duration.

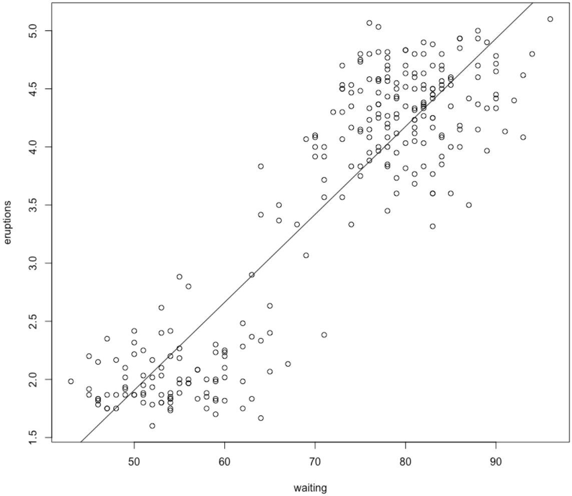
Example

The sample statistics for these data are

	`waiting`	`eruptions`
mean	$\bar{x} = 70.90$	$\bar{y} = 3.49$
sd	$s_x = 13.60$	$s_y = 1.14$
		$R = 0.90$

Find the linear regression line and interpret the parameter estimates.

Example



Hypothesis Testing in Linear Regression

Whenever we estimate a parameter, we want to use a hypothesis test to think about our confidence in that estimate.

- For β_i $(i = 0, 1)$

$$\begin{aligned} H_0: &\quad \beta_i = 0 \\ H_A: &\quad \beta_i \neq 0 \end{aligned}$$

- We will do this using a one-sample t-test.

Example

If we use `R` to get the coefficients for our `faithful` data, we get

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	-1.874016	0.160143	-11.70	<2e-16
waiting	0.075628	0.002219	34.09	<2e-16

What does this tell us about our parameters?

Extrapolation

- When we make predictions, we simply plug in values of x to estimate values of y.
- However, this has limitations!
- We don't know how the data outside of our limited window will behave.

Extrapolation

Applying a model estimate for values outside of the data's range for x is called **extrapolation**.

- The linear model is only an approximation.
- We don't know anything about the relationship outside of the scope of our data.
- Extrapolation assumes that the linear relationship holds in places where it has not been analyzed.

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 6^{th} 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.

Stephen Colbert
April 6th, 2010[12]

Example

- In this data, waiting times range from 43 minutes to 96 minutes.
- Let's predict
  - eruption time for a 50 minute wait.
  - eruption time for a 10 minute wait.

Using R^2 to Describe Strength of Fit

We've evaluated the strength of a linear relationship between two variables using the correlation coefficient R.

However, it is also common to use R^2. This helps describe how closely the data cluster around a linear fit.

Using R^2 to Describe Strength of Fit

Suppose $R^2 = 0.62$ for a linear model. Then we would say

- About 62% of the data's variability is accounted for using the linear model.

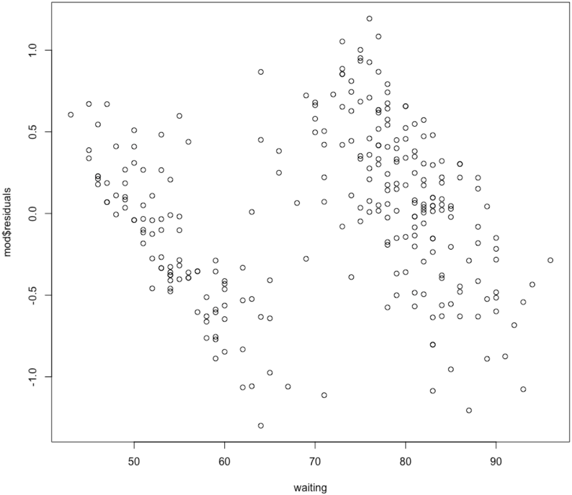
And yes, R^2 is the square of the correlation coefficient R!

Example

```
> summary(lm(eruptions~waiting))

Call:
lm(formula = eruptions ~ waiting)

Residuals:
     Min       1Q   Median       3Q      Max
-1.29917 -0.37689  0.03508  0.34909  1.19329

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
waiting      0.075628   0.002219   34.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4965 on 270 degrees of freedom
Multiple R-squared:  0.8115,	Adjusted R-squared:  0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

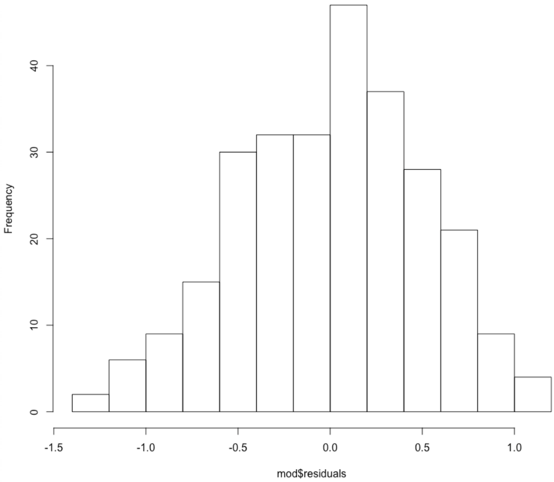
Interpret the R^2 value for this model.
What else can we learn from the R output?

Regression Example



This is the residual plot for the geyser regression. Do you see any problems?

Regression Example



This is a histogram of the residuals. Do they look normally distributed?