Regression and Difference of Two Proportions

August 28, 2019

Regression 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

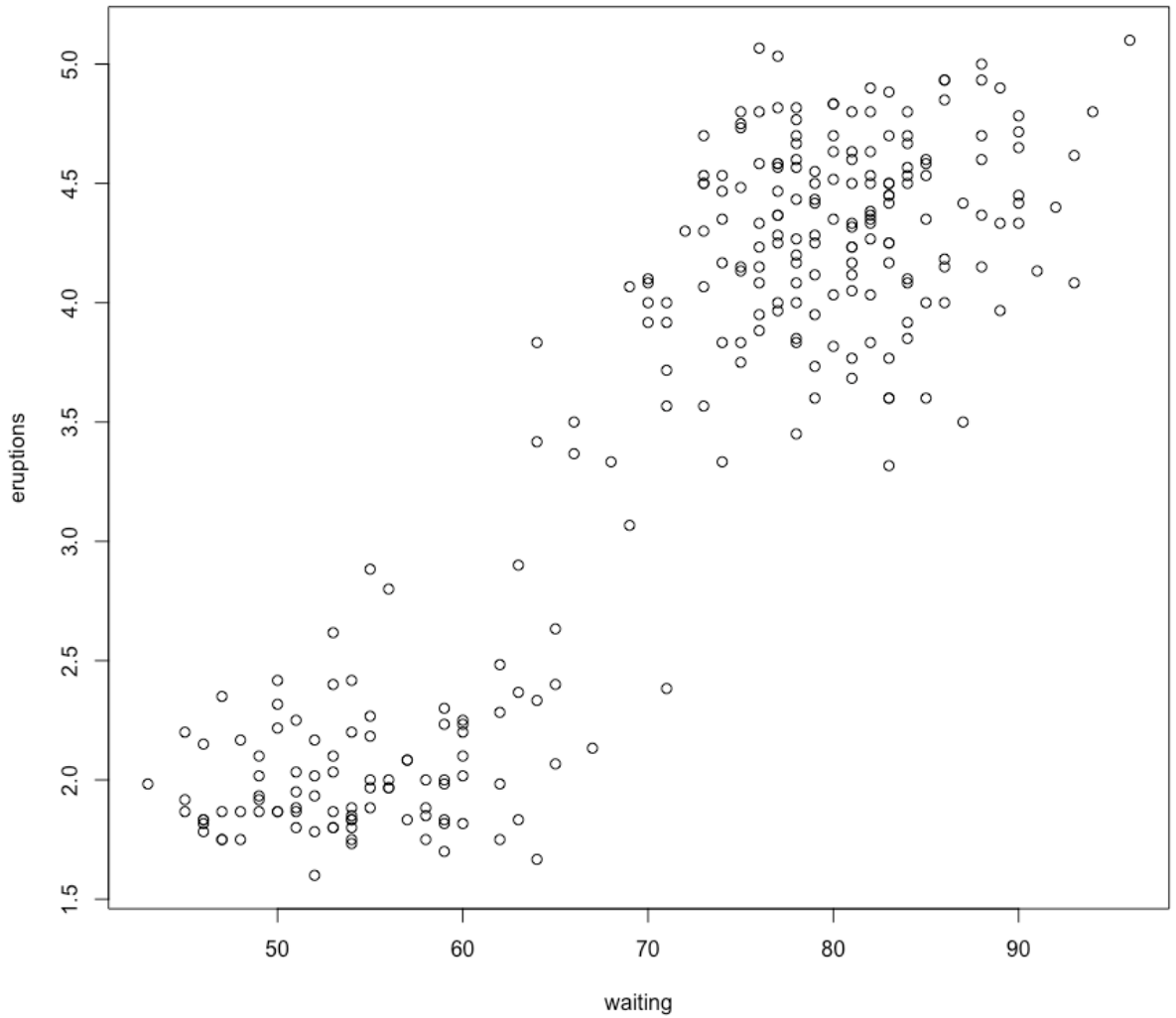
Each is measured in minutes.

Regression Example

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

- `eruptions` will be the response variable.
- `waiting` will be the predictor variable.

Regression Example



Regression Example

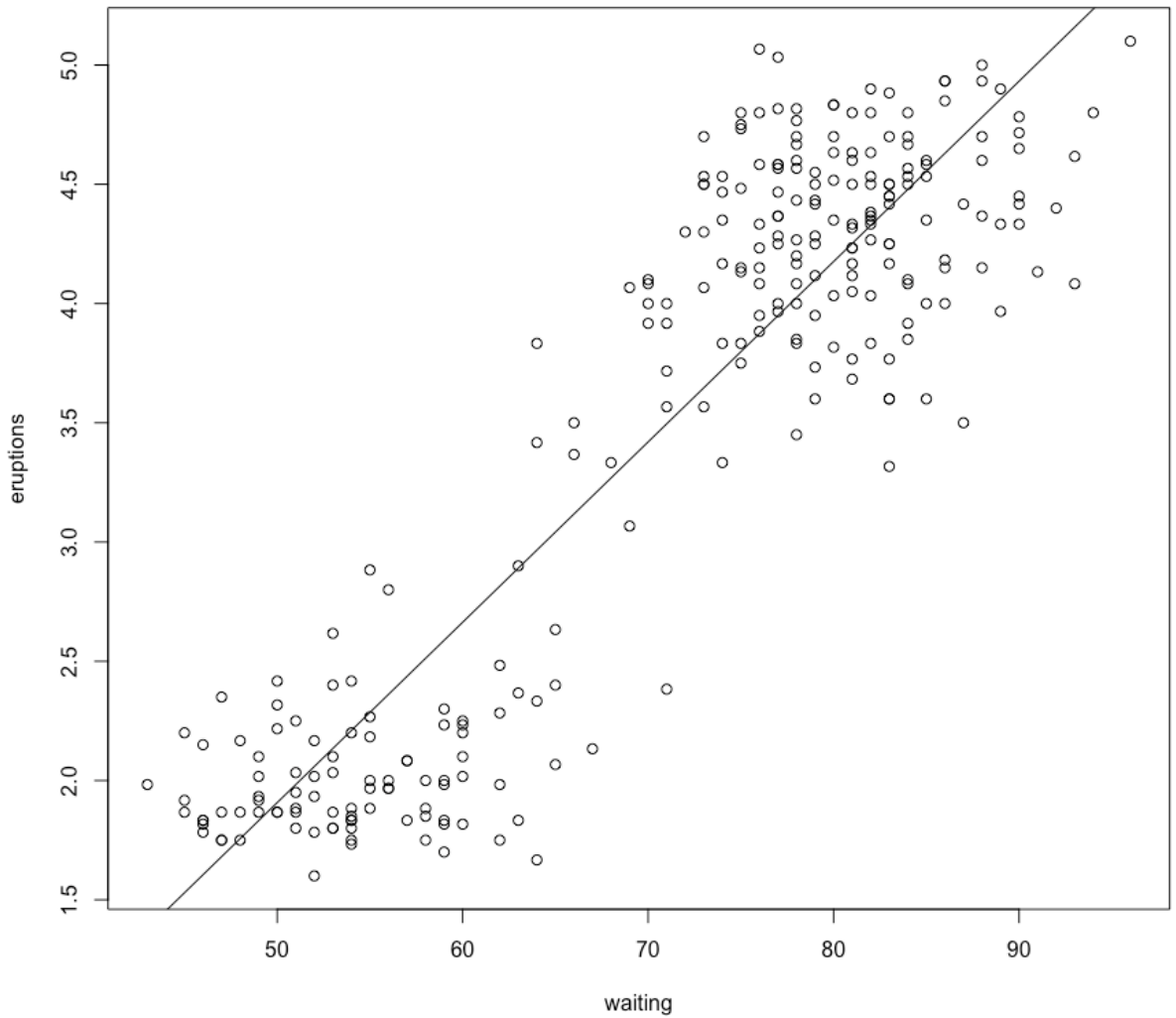
Using `R`, the estimated regression line for

$$\texttt{eruptions} = \beta_0 + \beta_1 \texttt{waiting} + \epsilon$$

is found to be

$$\hat{y} = -1.8740 + 0.0756x$$

Regression Example



Regression 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.

Regression Example

For `waiting` $= x = 50$,

$$\begin{aligned}\hat{y} &= -1.8740 + 0.0756x \\ &= -1.8740 + 0.0756 \times 50 \\ &= 1.906\end{aligned}$$

So for a wait time of 50 minutes, the predicted average eruption time is 1.906 minutes.

Regression Example

For `waiting` $= x = 10$,

$$\begin{aligned}\hat{y} &= -1.8740 + 0.0756x \\ &= -1.8740 + 0.0756 \times 10 \\ &= -1.118\end{aligned}$$

So for a wait time of 10 minutes, the predicted average eruption time is -1.118 minutes.

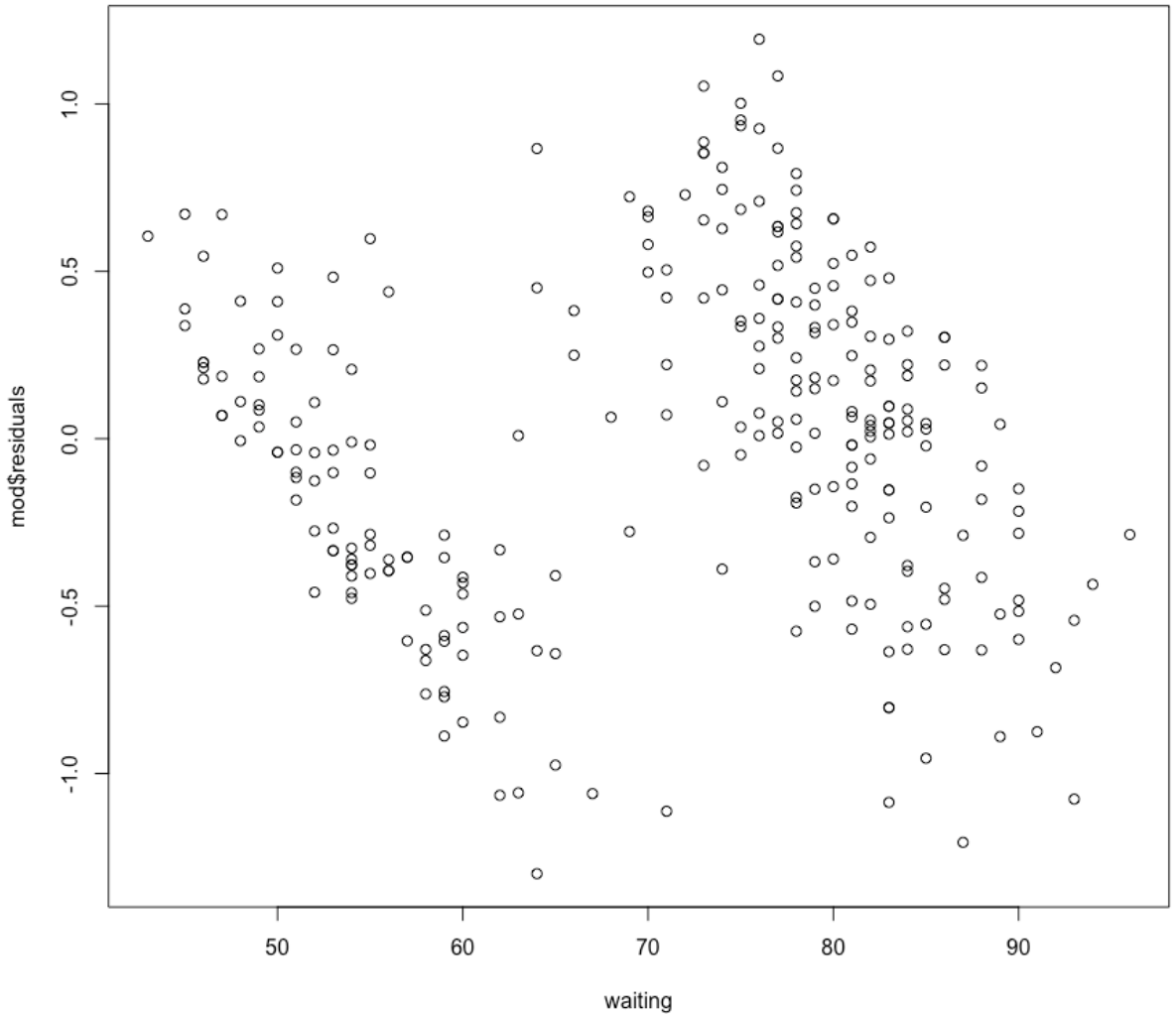
Regression Example

But a predicted average eruption time of -1.118 minutes

1. doesn't make sense.
2. is an extrapolation!

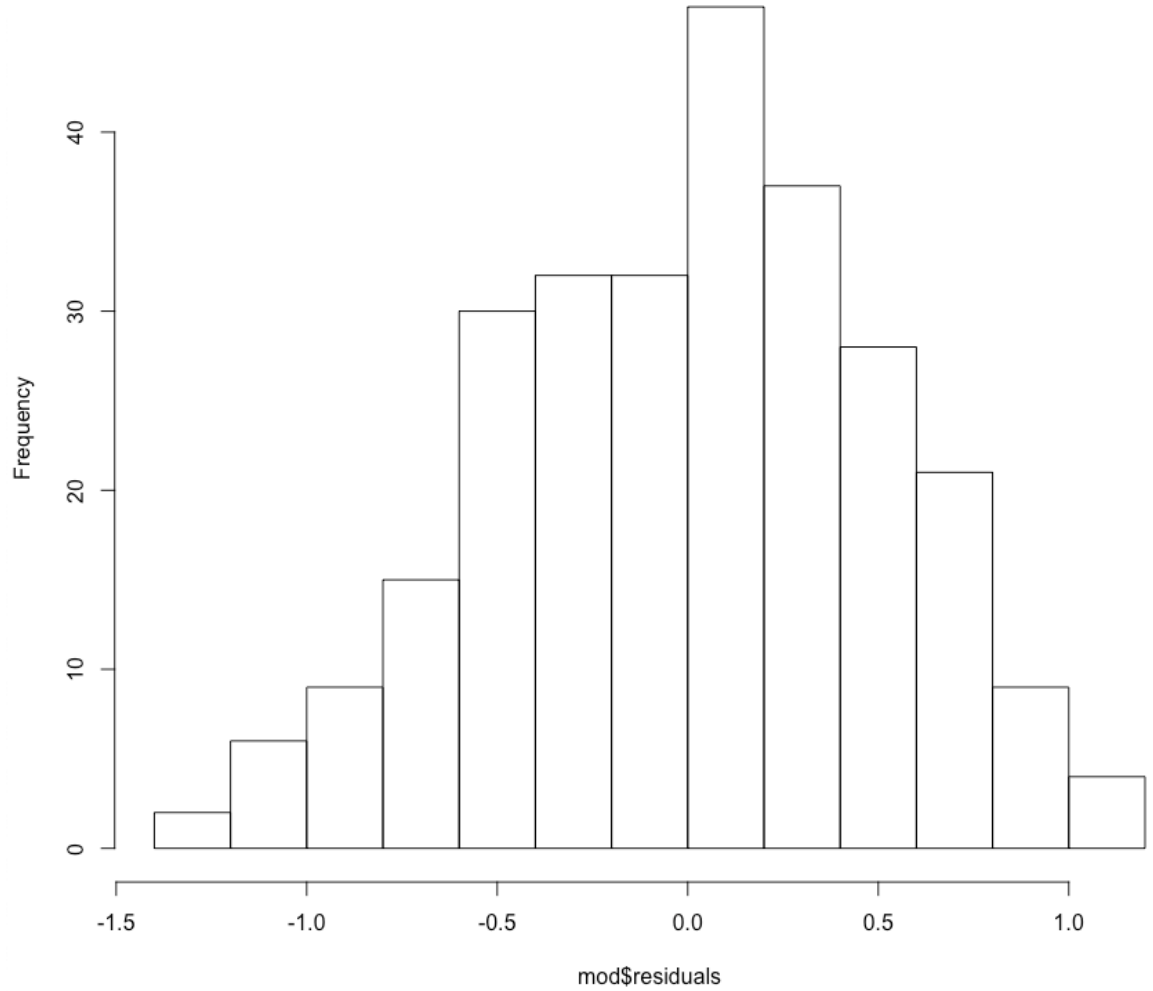
We do not want to make this prediction.

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?

Regression Example

Asking R for a summary of the regression model, we get the following:

```
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
```

Let's pick this apart piece by piece.

Regression Example

```
Call:
lm(formula = eruptions ~ waiting)

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

- The first line shows the command used in `R` to run this regression model.
- The `Residuals` item shows a quartile-based summary of our residuals.

Regression Example

```
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
```

The `F-statistic` and `p-value` give information about the model overall.

- These are based on an F-distribution.
- The null hypothesis is that all of our model parameters are 0 (the model gives us no good info).
- Since p-value$< 2.2 \times 10^{-16} < \alpha = 0.05$, at least one of the parameters is nonzero (the model is useful).

Regression Example

```
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
```

- `Multiple R-squared` is our squared correlation coefficient R^2.
- Ignore the adjusted R-squared and residual standard error for now.

Regression Example

```
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
```

Finally, the `Coefficients` section gives us several pieces of information:

1. `Estimate` shows the estimated parameters for each value.
2. `Std. Error` gives the standard error for each parameter estimate.
3. The `t valuess` are the test statistics for each parameter estiamte.
4. Finally, `Pr(>|t|)` are the p-values for each parameter estimate.

Regression Example

The hypothesis test for each regression coefficient has hypotheses

$$H_0 : \beta_i = 0$$
$$H_A : \beta_i \neq 0$$

where $i = 0$ for the intercept and $i = 1$ for the slope.

Regression Example

```
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
```

1. $p - value < 2 \times 10^{-16}$ for b_0 so we can conclude that the intercept is nonzero.
2. $p - value < 2 \times 10^{-16}$ for b_1 so we conclude that the intercept is also nonzero.
3. This means that the intercept and slope both provide useful information when predicting values of $y =$ `eruptions`.

Difference of Two Proportions

- We will extend the methods for hypothesis tests for p to methods for $p_1 - p_2$.
- This is the difference of proportions for two different groups or populations.
- The point estimate for $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$.
- We will develop a framework for use of the normal distribution and a new standard error formula.

Conditions for Normality

$\hat{p}_1 - \hat{p}_2$ may be modeled using a normal distribution when

- The data are independent within and between groups.
  - This should hold if the data from from a randomized experiment or from two independent random samples.
- Success-failure condition holds for both groups.

$$n_1 p_1 \geq 10 \quad \text{and} \quad n_1(1 - p_1) \geq 10$$

and

$$n_2 p_2 \geq 10 \quad \text{and} \quad n_2(1 - p_2) \geq 10$$

Standard Error

When the normality conditions hold, the standard error of $\hat{p}_1 - \hat{p}_2$ is

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

where p_1 and p_2 are the proportions and n_1 and n_2 are their respective sample sizes.

Confidence Intervals

We can again use our generic confidence interval formula

$$\text{point estimate} \pm \text{critical value} \times SE$$

now as

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Confidence Intervals

The intervals are interpreted as before. E.g.,:

One can be 95% confident that the true difference in proportions is between lower bound and upper bound.

Hypothesis Tests: Example

- A 30-year study was conducted with nearly 90,000 female participants.
- During a 5-year screening period, each woman was randomized to one of two groups: regular mammograms or regular non-mammogram breast cancer exams.
- No intervention was made during the following 25 years of the study, and we'll consider death resulting from breast cancer over the full 30-year period.

Hypothesis Tests: Example

Over the 30-year period,

- of the 44,925 women receiving mammograms, 500 died from breast cancer.
- of the 44,910 women receiving other cancer detection exams, 505 died from breast cancer.

Create a contingency table for these data.

Hypothesis Tests: Example

Set up the hypotheses for these data.

Special Case

When H_0: $p_1 = p_2$, we use a special **pooled proportion** to check the success-failure condition:

$$\hat{p}_{pooled} = \frac{\text{number of "yes"}}{\text{total number of cases}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

Note that this is usually the null hypothesis used in tests for two proportions.

Hypothesis Tests: Example

Let's calculate $\hat{p}_{pooled}$ or our mammograms example.

We will use this to check the success-failure condition.

Pooled Standard Error

When H_0: $p_1 = p_2$, the standard error is calculated as

$$SE_{pooled} = \sqrt{\frac{p_{pooled}(1 - p_{pooled})}{n_1} + \frac{p_{pooled}(1 - p_{pooled})}{n_2}}$$

Hypothesis Tests: Example

Let's find the point estimate and standard error for our mammograms example.

Test Statistic

As before, the test statistic is calculated as

$$ts = z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{(\hat{p}_1 - \hat{p}_2) - (\text{null value})}{SE}$$

Hypothesis Tests: Example

For our mammograms example, the null value is 0, so

$$ts = z = \frac{(\hat{p}_1 - \hat{p}_2)}{SE}$$

The critical value is $z_{\alpha/2}$. At the 0.05 level of significance, $z_{0.025} = 1.96$.

Hypothesis Tests: Example

Since $|z_{0.025}| = 1.96 > |z| = |-0.17| = 0.17$,

- we fail to reject the null hypothesis.
- there is insufficient evidence to suggest that mammograms are either helpful or harmful.