Simple Linear Regression

STA 198: Introduction to Health Data Science

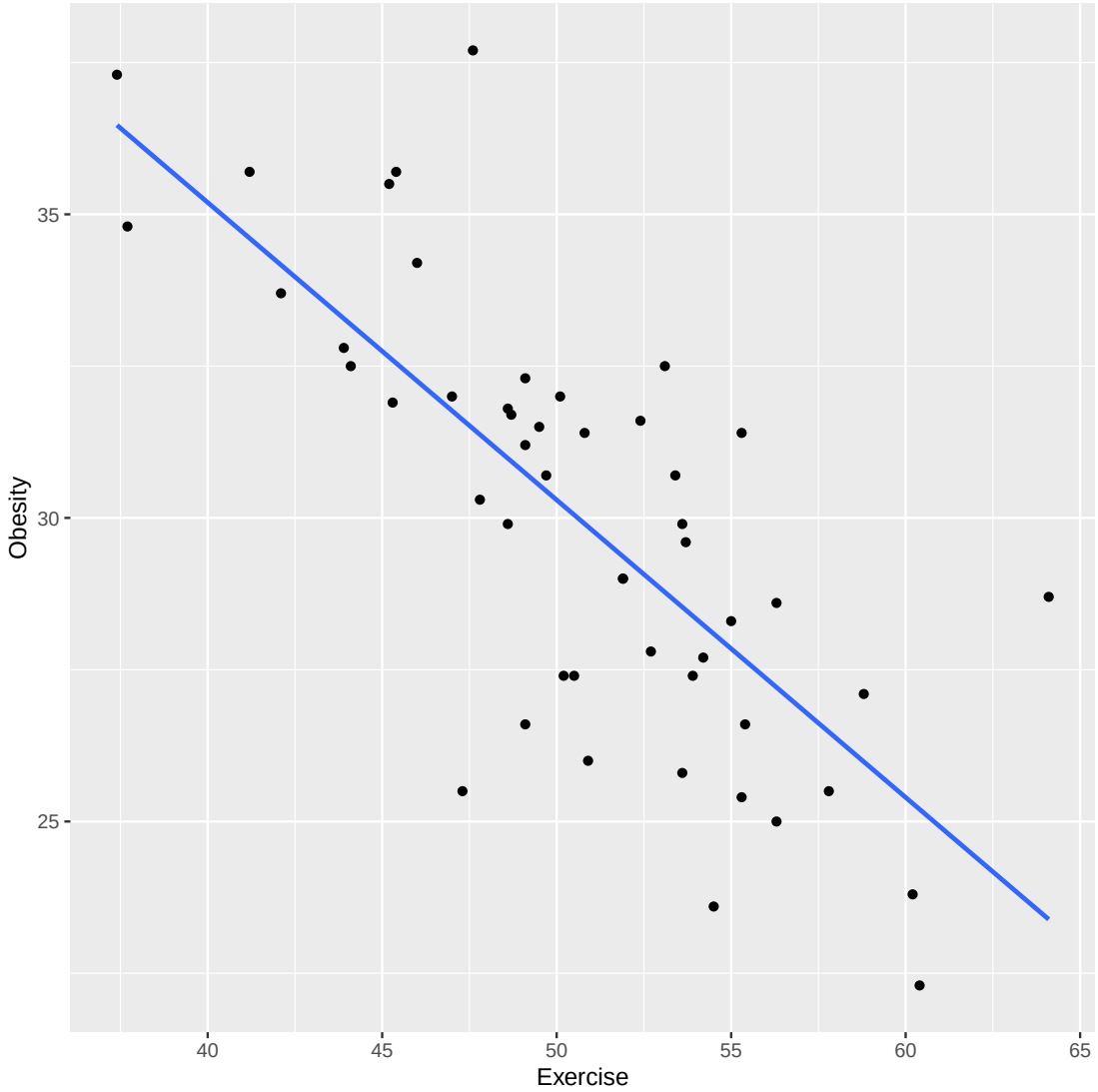
Yue Jiang

July 10, 2020

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

Back to Lab 01...



Correlation and regression

- Correlation measures the direction and strength of linear relationships
- Regression can be used to further quantify these relationships
- A regression line summarizes the relationship between explanatory or predictor variables (e.g., Exercise %, X) and response or outcome variables (e.g., Obesity % or Y)
- The fitted regression line can be used to predict the outcome for a given set of predictor values
- Our predictions do have error, called residuals
- The least-squares regression line minimizes the sum of the squared error

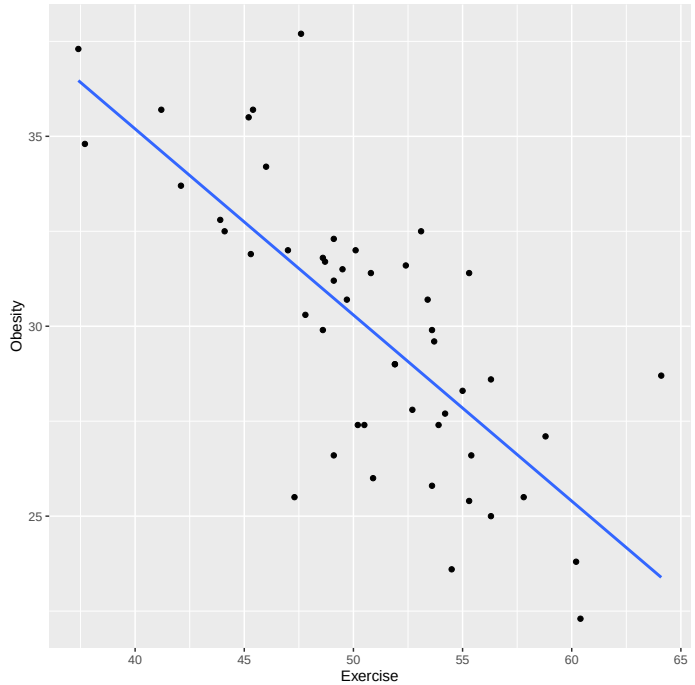
Regression

In regression, our we use one variable (or more) to try to predict values of another. In simple linear regression, one variable (Y) is the *response* or *outcome* or *dependent variable* and the other (X) is the *predictor* or *explanatory variable* or *independent variable*.

This distinction is critical. The regression of Y on X is not equal to the regression of X on Y.

The regression of Y on X can be used to predict Y based on fixed values of X.

Back to Lab 01...



- The regression line relating exercise and obesity percentages by state is shown
- Points represent actual data values
- Note the line is not a perfect fit

The model

The model is given by $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where

- y_i is the outcome (dependent variable) of interest
- β_0 is the intercept parameter
- β_1 is the slope parameter
- x_i is a predictor variable
- ϵ_i is the error (like β_0 and β_1, it is not observed)

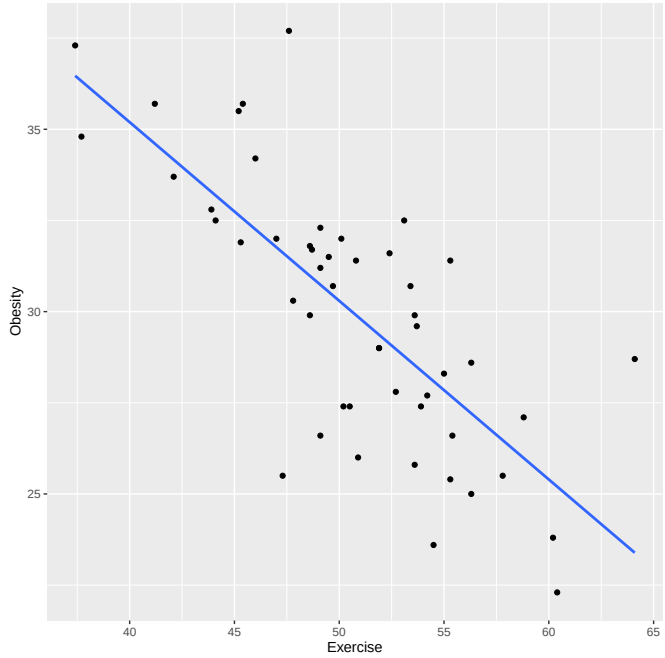
The i is an index of observations. So, each observation of the outcome is related to an observation of that individual's value of the predictor.

The least squares model

- Fitted line equation:
  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Residual:
  $\hat{\epsilon}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) = y_i - \hat{y}_i$
- Least squares selects $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the *error sum of squares*

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

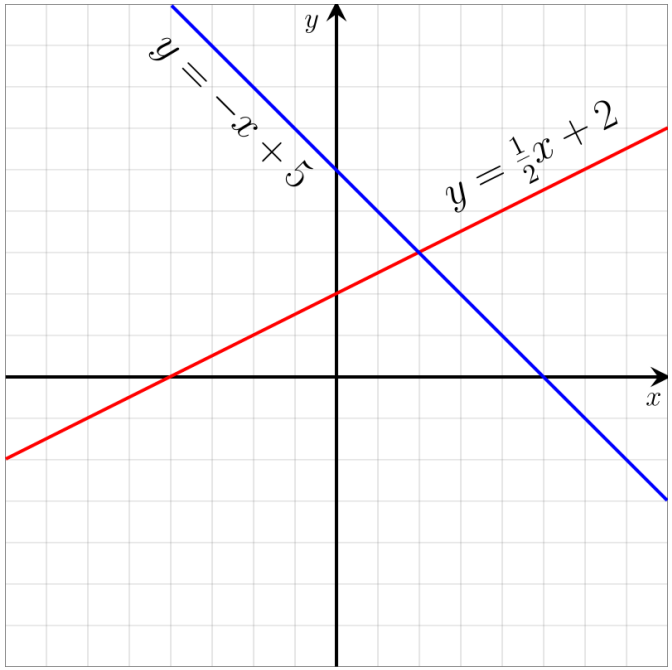
is minimized



Dots are ordered pairs (x_i, y_i)

Model: $y_i = \beta_0 + \beta_1 x_1 + \epsilon_i$

Review: slope-intercept form of a line



Recall the slope-intercept form of a line, $y = mx + b$.

For instance, in the red equation, $m = \frac{1}{2}$ and $b = 2$. In the blue equation, $m = -1$ and $b = 5$.

Review: slope-intercept form of a line

- b is the y-intercept, or where the line crosses the y-axis. It is the predicted value of y when $x = 0$
- m is the slope, which tells us the predicted increase in y when x changes by 1 unit
- For the red line on the previous slide ($y = \frac{1}{2}x + 2$), what is our predicted y when $x = 2$?
- In statistics, we often represent the slope with β_1 and the intercept with β_0, and these values are usually not known but must be estimated.

Model output

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 54.79087    3.24772  16.871  < 2e-16 ***
Exercise    -0.48991    0.06365  -7.697 6.34e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.455 on 48 degrees of freedom
Multiple R-squared:  0.5524,    Adjusted R-squared:  0.5431
F-statistic: 59.25 on 1 and 48 DF,  p-value: 6.336e-10
```

The equation for the fitted line is $\hat{y}_i = 54.79 - 0.49x_i$.

- The intercept is the estimated mean outcome when the predictor is zero (does this always make sense?)
- The slope is the expected change in the outcome corresponding to a one-unit change in the predictor

Note that we often use hats to denote estimates. For instance, $\hat{\beta}_1$ is our estimate of the slope β_1

How might we interpret these coefficient estimates in our model?

Model assumptions

Assumptions of the model $y_i = \beta_0 + \beta_1 x_1 + \epsilon_i$ include:

- The outcomes y_i are *independent*. This is violated if our study contains repeated outcome measures on an individual, if siblings are enrolled, etc.
- A linear relationship holds (though we can relax this to some extent)
- For a specified value of x, which is measured without error, $y_i \sim N(\beta_0 + \beta 1 x_i, \sigma^2)$. Note that this implies $\epsilon_i \sim N(0, \sigma^2)$.
- The variance σ^2 is constant across all values of x; this is called homogeneity of variance

Hypothesis testing

- The primary hypothesis test of interest is usually $H_0 : \beta_1 = 0$ (no association between exercise and obesity percentage) versus $H_1 : \beta_1 \neq 0$.
- Under the null hypothesis, we can use a t-test to test this.

Given our t-statistic of -7.7 and p-value of $<$0.001, what might we conclude in our data?

Some additional model output...

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 54.79087    3.24772  16.871  < 2e-16 ***
Exercise    -0.48991    0.06365  -7.697 6.34e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.455 on 48 degrees of freedom
Multiple R-squared:  0.5524,    Adjusted R-squared:  0.5431
F-statistic: 59.25 on 1 and 48 DF,  p-value: 6.336e-10
```

The equation for the fitted line is $\hat{y}_i = 54.79 - 0.49x_i$.

- The R^2 indicates that exercise % explains 55% of the variance of obesity %. R^2 tells us about how much of the variability in obesity % is explained by exercise; as R^2 gets closer to 1, the points get tighter around the regression line

Predicted values

The regression equation can be used to get predicted means at any value of x. $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, where we call $\hat{y}$ the predicted value of y at a given value of x (what happened to the error term?). For the CDC data that predicts a state's obesity percentage by its adequate exercise percentage, we have

$$\hat{y}_i = 54.79 - 0.49 x_i.$$

So, the predicted mean obesity percentage for a state where 50% of residents get adequate exercise is

$$\hat{y}_i = 54.79 - 0.49(50).$$

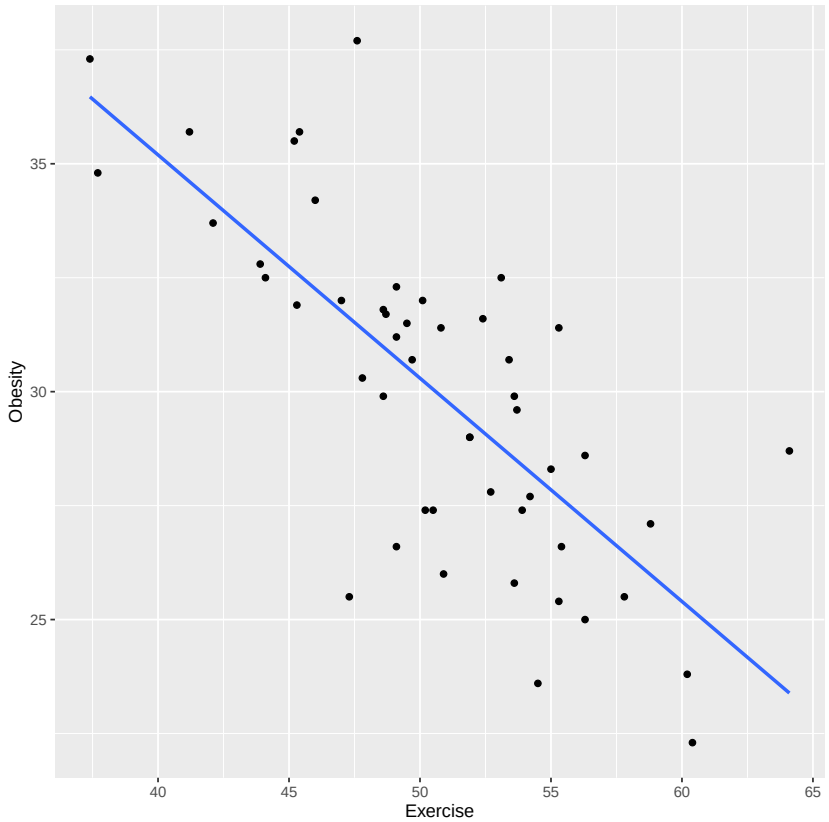
We would predict an obesity percentage of approximately 30%. Note the importance of getting the units correct.

Regression diagnostics: independence

We just have to think through this one, unfortunately.

Regression diagnostics: linearity

To judge linearity, look at our plot of the regression line superimposed on the data points. Is the line generally consistent with the point locations, or is it missing a nonlinear pattern?



Regression diagnostics: normality of errors

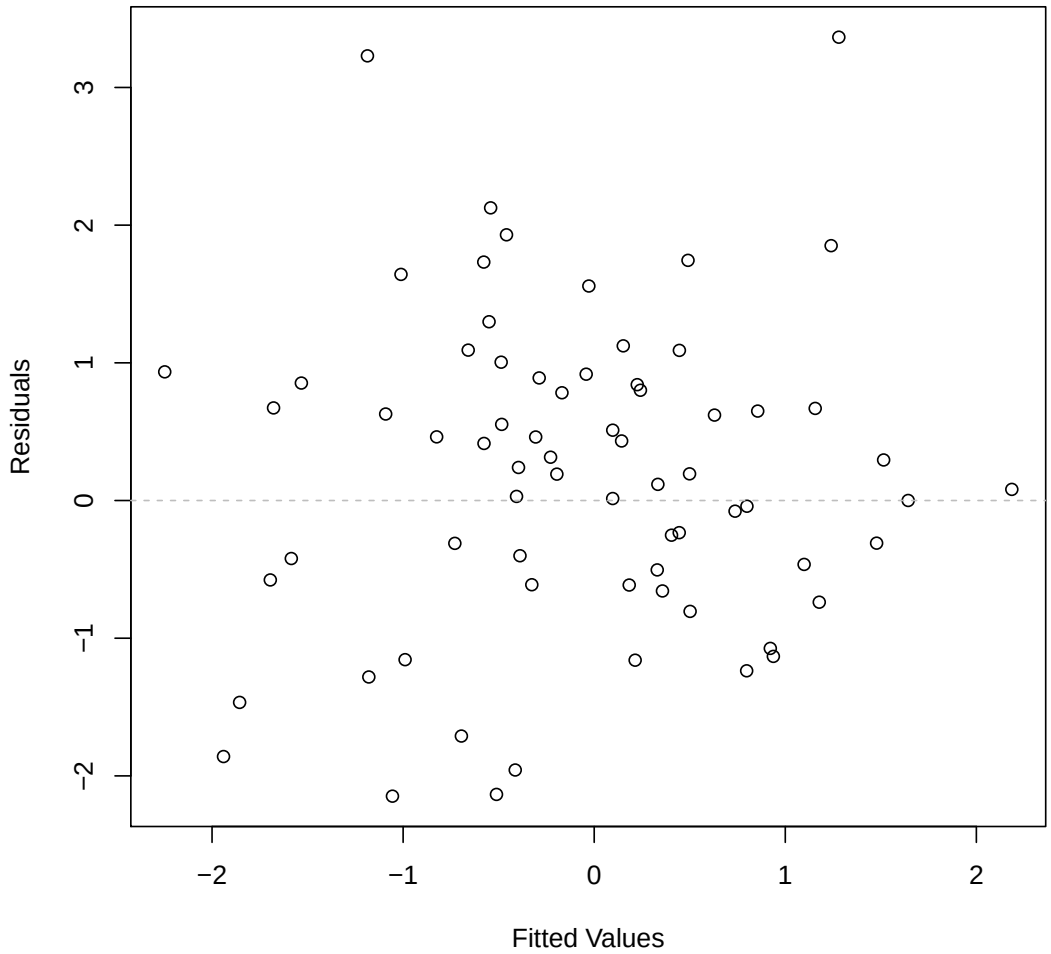
Use a histogram of the residuals to check normality. There are more advanced methods and plots, but we won't talk about them here (if you're interested, come to office hours!)

Regression diagnostics: equal variances

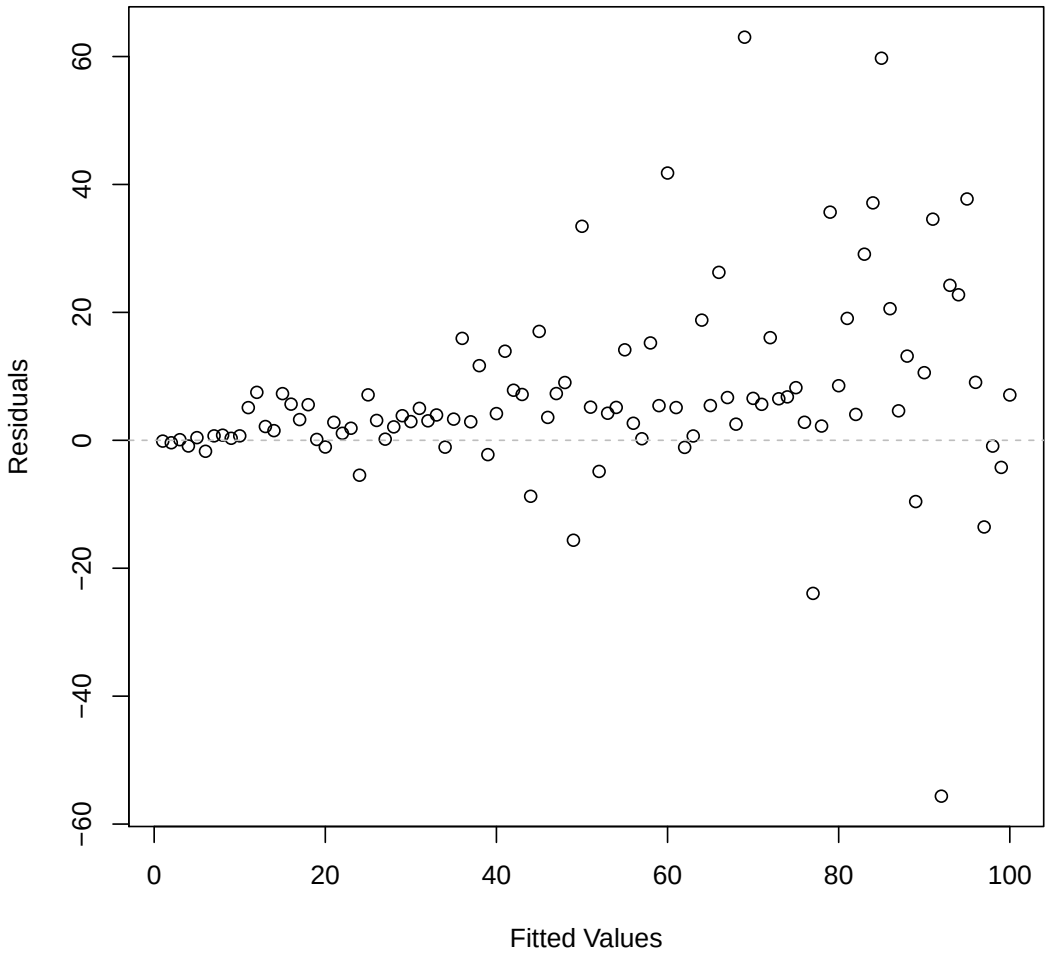
We should always check assumptions of regression to see whether the method is valid in our setting. To check equal variances, we can use a plot of the residuals ($\hat{\epsilon}_i$) by the predicted (or fitted) values ($\hat{y}_i$)

We expect to see evenly-spaced dots along the y axis. Patterns or trends are evidence something is wrong.

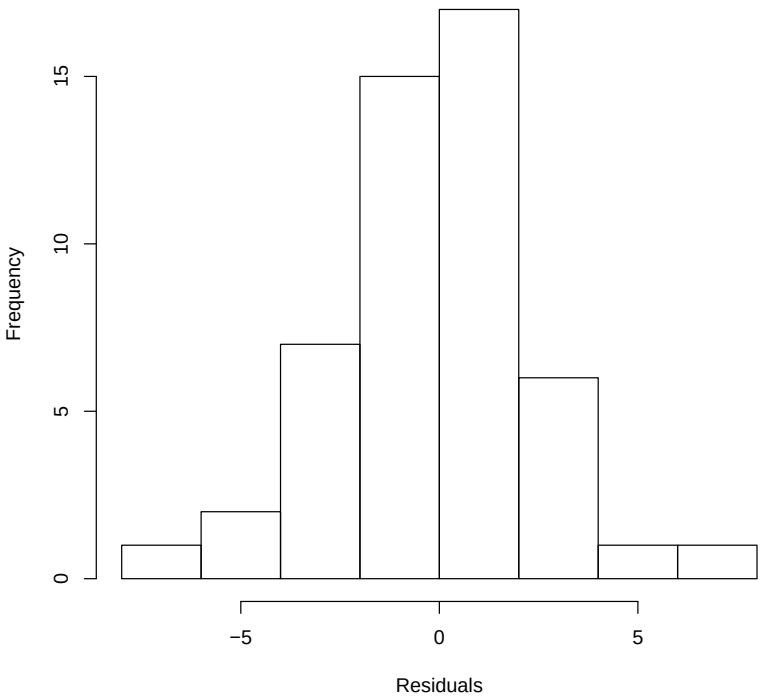
Equal variances

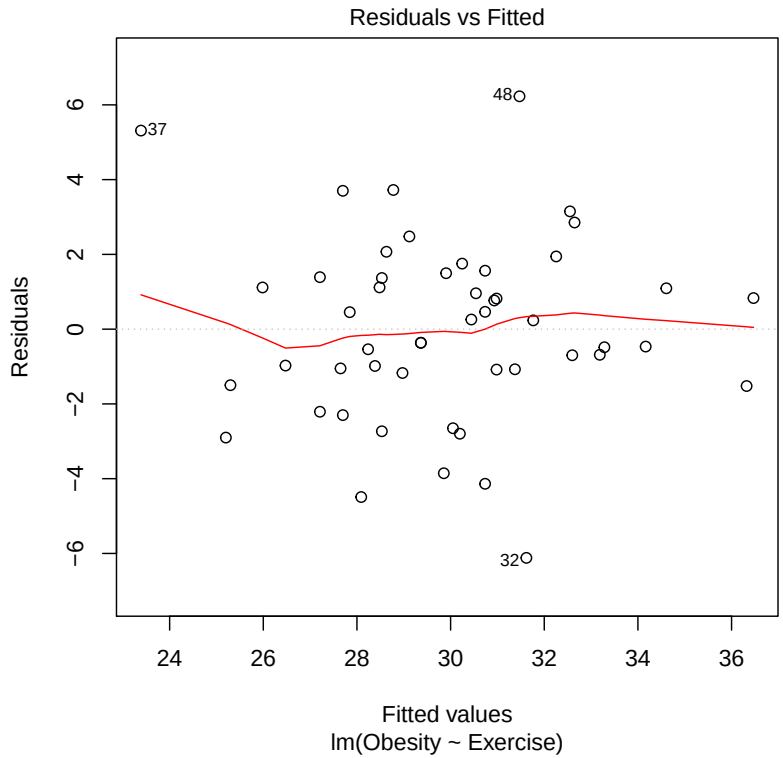


Unequal variances



Let's check assumptions for our model...





Assumptions summary

1. Independent observations
2. Linear relationship between predictor and response (can be relaxed using polynomials or splines)
3. Conditional on covariates, the response follows a normal distribution (normally distributed residuals)
4. Equal variances (homogeneity of variance in the residuals)