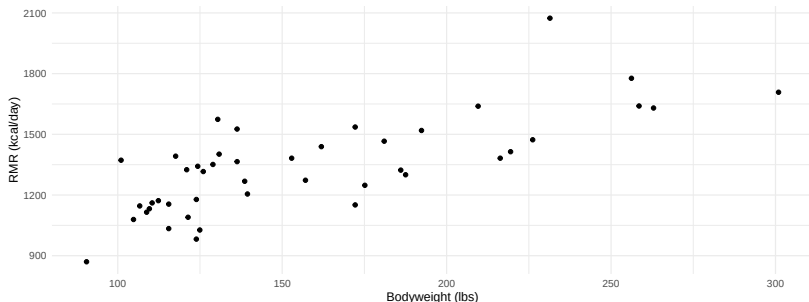


Simple Linear Regression

Ryan Miller

Introduction

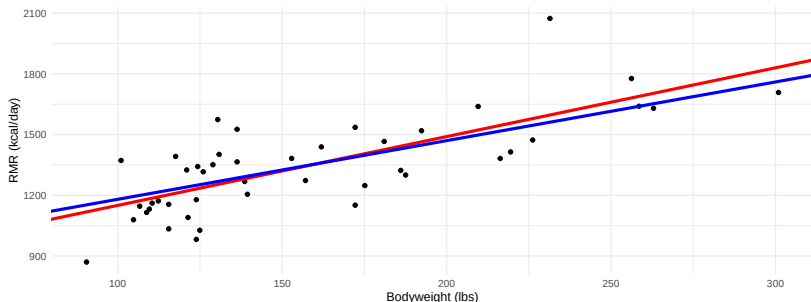
Shown below are the resting metabolic rate (RMR) and bodyweight of $n = 44$ adult women:



Clearly there is an association between these two variables, but could we use a woman's bodyweight to predict their resting metabolic rate?

Prediction Lines

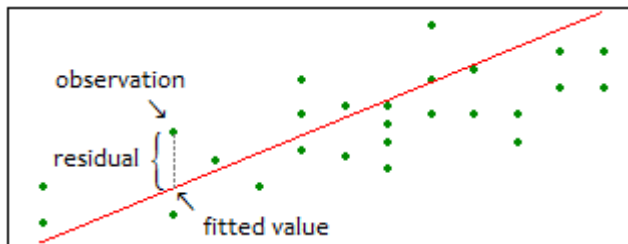
One approach is to *fit a line* to the data:



But there are infinitely many lines we could use (two reasonable ones are shown here). Which line should we use?

Least Squares Regression

Ideally, we want the line that is closest to data, and we can measure “closeness” using **residuals**:



The **least squares regression** line is the one that produces the minimum sum of squared residuals, or the lowest amount of squared error.

Least Squares Regression

Summarizing:

- ▶ We want to predict an outcome using an explanatory variable using a straight line as our model
- ▶ All lines have the form: $Y = \text{intercept} + \text{slope} \cdot X$
- ▶ Different lines produce different residuals, and thus different amounts of prediction error
- ▶ We want numerical values of the slope and y-intercept that minimize the squared residuals for the sample data (the smallest total prediction error)

Statistical Models

In a regression analysis, there are two distinct types of models we might reference:

- ▶ **Population model:** a proposed theoretical model stating there is a linear relationship between X and Y with some error
 - ▶ $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- ▶ **Fitted model:** an estimate of the population model found using the sample data
 - ▶ $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

We can form hypotheses about the components of the population model (ie: $H_0 : \beta_1 = 0$) and assess them using the fitted model (ie: use $\hat{\beta}_1 = 3.1$ to find a p -value measuring the evidence against H_0).

Practice

In the resting metabolism example, the least squares regression line has a slope of 3.4 and an intercept of 811.2.

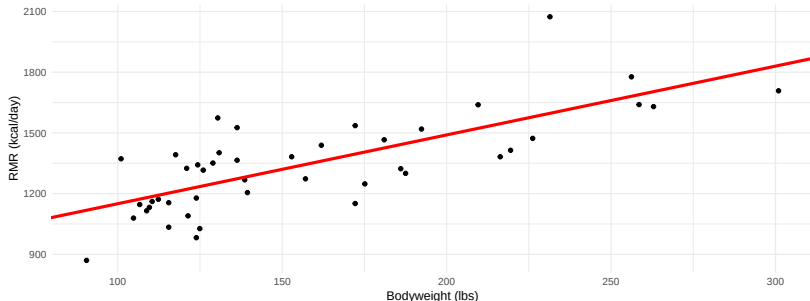
1. Write out the population and fitted model for this example.
2. A woman weighs 150 pounds, what does the fitted model predict for this woman's RMR?
3. Suppose the 150 pound woman had an observed RMR of 1500 kcal/day. Did the model overestimate or underestimate her RMR? What is her *residual*?

Interpretation

- ▶ The regression line predicts the *average value of the response variable* at a given value of X
 - ▶ Thus, a model prediction of $\hat{y} = 1321.2$ kcal/day when $X = 150$ lbs doesn't imply we believe that every 150 pound woman will have an RMR of 1321.2 kcal/day.
 - ▶ Instead, women weighing 150 pounds are expected to have an *average RMR* of 1321.2 kcal/day.
- ▶ Similarly, this means that the estimated slope, $\hat{\beta}_1$, tells about the *expected change in the average value of Y* for a certain change in X
 - ▶ $\hat{\beta}_1 = 3.4$ tells us that, on average, we expect a 1 pound increase bodyweight to correspond with a 3.4 kcal/day increase in RMR
 - ▶ We *do not* expect every individual's RMR to increase by exactly 3.4 kcal/day for each additional pound of bodyweight

Model Fit

Below is the fitted least squares regression line for the RMR example:



Bodyweight seems to be predictive of RMR, but how much information does it actually provide?

Model Fit

The **coefficient of determination**, R^2 , is a popular measure of well a model explains an outcome

$$R^2 = \frac{\text{variation explained by model}}{\text{total variation in } Y} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

- ▶ Larger values of R^2 reflect stronger adherence to the linear trend
- ▶ If $R^2 = 0$, the regression line explains none of the variability in Y
- ▶ If $R^2 = 1$, the regression line perfectly predicts every observed value of Y

The regression line we've been looking at was built using a sample of $n = 44$ women. If we had collected a different sample:

1. The slope and intercept would likely be somewhat different
2. The R^2 of the line would likely be somewhat different

So, because sample to sample variation, we might wonder whether our data provide compelling statistical evidence of a linear relationship between body weight and RMR. . .

Consider $H_0 : \beta_1 = 0$

1. If H_0 were true, what would you expect the *sample data* to look like?
2. If H_0 were true, what would you expect the *fitted regression line* to look like?
3. What would we use as evidence against H_0 ?

If the null hypothesis were true, the test statistic:

$$T = \frac{\hat{\beta}_1 - 0}{SE}$$

follows a t -distribution with $n - 2$ degrees of freedom.

- ▶ As we've seen, larger values of T indicate more evidence against H_0 (smaller p -value)
 - ▶ Thus, the p -value will be smaller with a larger magnitude slope (bigger $\hat{\beta}_1$), or a smaller SE (less sampling variability, primarily driven by a larger sample size)

Model Inference in R

The code below shows *t*-test results for both the slope and intercept:

```
rmr <- read.csv("https://remiller1450.github.io/data/RMR.csv")  
fitted_model = lm(rate_kcal ~ weight_lbs, data = rmr)  
summary(fitted_model)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	811.22667	76.9755034	10.538764	2.288384e-13
## weight_lbs	3.36168	0.4655228	7.221301	7.025380e-09

Model Inference in R

- ▶ Could you calculate the T -value of 7.22 using other parts of the R output? If so, how?
- ▶ What do you conclude from the p -value for `weight_lbs`?
What about the p -value for the intercept?

Connections Between Regression, Correlation, and R^2

- ▶ In simple linear regression, $R^2 = r^2$
 - ▶ So, $\hat{\beta}_1 = 0 \iff R^2 = 0 \iff r = 0$
- ▶ The slope, correlation coefficient, and R^2 are all describing the same underlying linear relationship
 - ▶ As a consequence, testing $H_0 : \beta = 0$ is equivalent to testing $H_0 : r = 0$ or $H_0 : R^2 = 0$

Practice

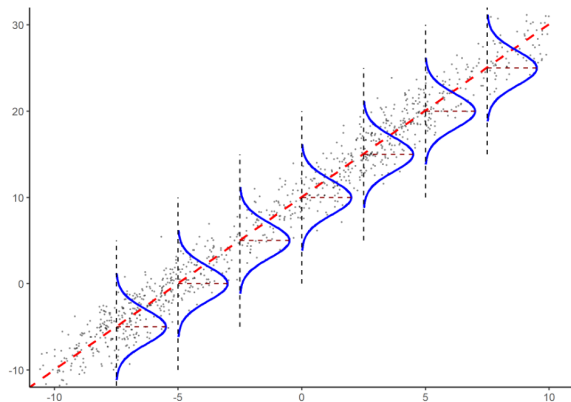
For the Grinnell Peers dataset (provided below), fit a linear regression model using `Adm_Rate` to predict `Med_Grad_Earnings_4Y` and use the `summary()` function to print a summary of the fitted model.

1. Interpret the slope coefficient of the fitted model. *Hint*: Does a 1-unit change in a proportion (ranging from 0-1) make sense? How should you adjust your interpretation?
2. Briefly explain what the T -test results for `Adm_Rate` tell us about the relationship between admissions rates and median graduate earnings among liberal arts colleges like Grinnell?

```
grinnell_peers = read.csv("https://remiller1450.github.io/data/grinnell_peers.csv")
```

Assumptions and Residual Plots

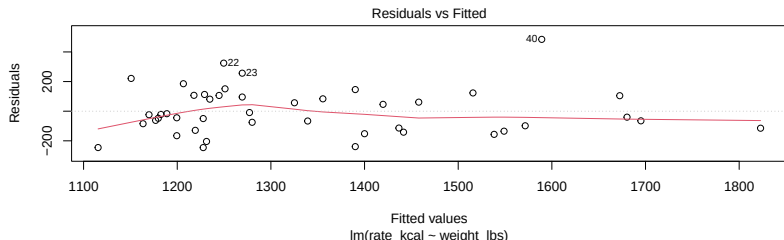
The T -test for β_1 was derived by assuming the model's errors are independent and follow an identical Normal distribution, regardless of the value of X :



Assumptions and Residual Plots

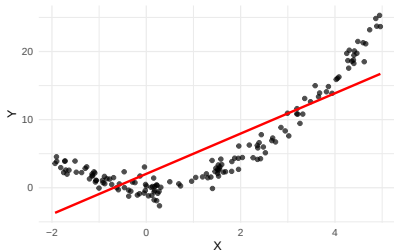
We can check these assumptions by plotting the residuals vs. predicted values. If the points look like random scatter with roughly constant spread, the assumptions are reasonable:

```
fitted_model = lm(rate_kcal ~ weight_lbs, data = rmr)
plot(fitted_model, which = 1)
```

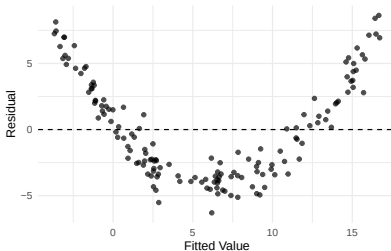


Assumptions and Residual Plots

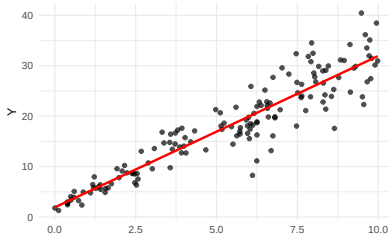
Nonlinear Data



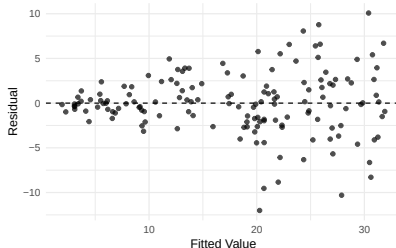
Residuals vs Fitted



Data with Changing Variability



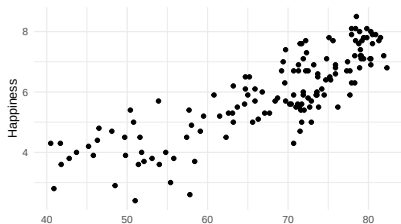
Residuals vs Fitted



Practice

1. Fit a linear regression model using LifeExpectancy to predict Happiness and check the conditions for statistical inference using the `plot()` function (see earlier example)
2. Now use GDPperCapita to predict Happiness and check the conditions for statistical inference. How do these residual plots relate to the scatterplots given below?

```
happy_planet = read.csv("https://remiller1450.github.io/data/HappyPlanet.csv")
```



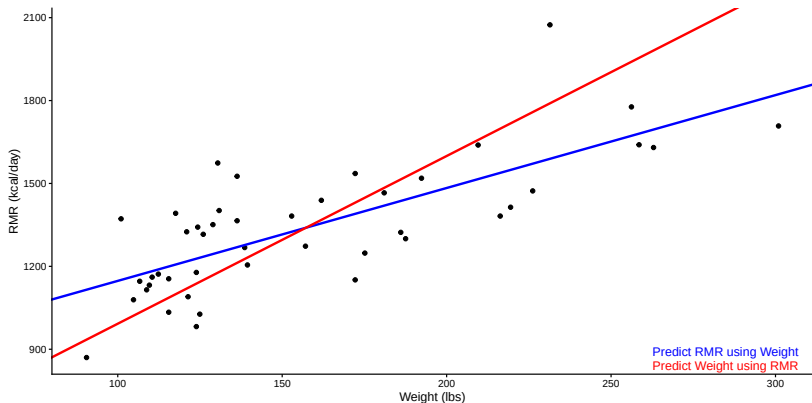
Common Pitfalls

In a regression analysis, several common mistakes are:

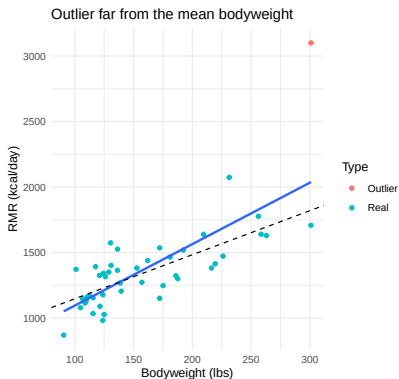
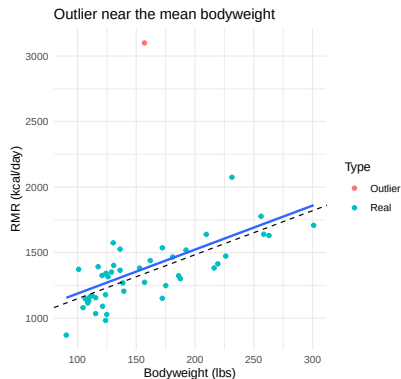
1. Reversing explanatory and response variables
2. Misunderstanding the impact of unusual/influential observations
3. Extrapolating beyond the range of the observed data

Regression Mistake #1

Notice you don't get the same line if you reverse the explanatory and response variables:



Regression Mistake #2



Outliers in the response variable only exert a disproportionate impact on the regression line if they are also far from the average value of the explanatory variable

Regression Mistake #3

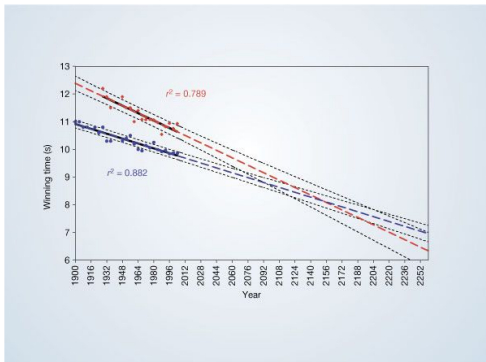
In 2004, an article was published in *Nature* titled “Momentous sprint at the 2156 Olympics”. The authors plotted the winning times of the men’s and women’s 100m dash in every Olympics, fitting separate regression lines to each. They found that the lines will intersect at the 2156 Olympics, here are a few media headlines:

- ▶ “Women ‘may outsprint men by 2156’ ” - BBC News
- ▶ “Data Trends Suggest Women will Outrun Men in 2156” - Scientific American
- ▶ “Women athletes will one day out-sprint men” - The Telegraph
- ▶ “Why women could be faster than men within 150 years” - The Guardian

Do you have any problems with these conclusions?

Extrapolation

Here's a figure from the original publication in Nature:

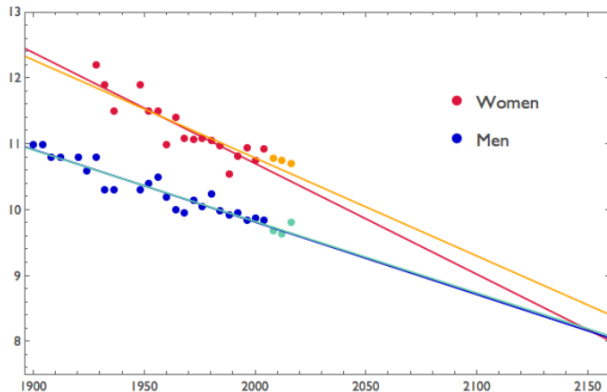


The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

Extrapolation

source: https://callingbullshit.org/case_studies/case_study_gender_gap_running.html

Since the *Nature* paper was published, we've had three additional Olympic games (yellow and green points below) and see how the model has performed.



Key Takeaways

- ▶ Regression uses an explanatory variable to predict the average value of a response variable.
 - ▶ Statisticians will frequently reference a population-level model, and a model fitted to the sample data
- ▶ The fitted least-squares line is found by minimizing the sum of squared residuals.
- ▶ Hypothesis tests about β_1 (the slope) tell us about the relationship between X and Y in the population, and are connected to both correlation (r) and R^2
 - ▶ We should check residual plots to make sure the p -value produced by this test is reliable