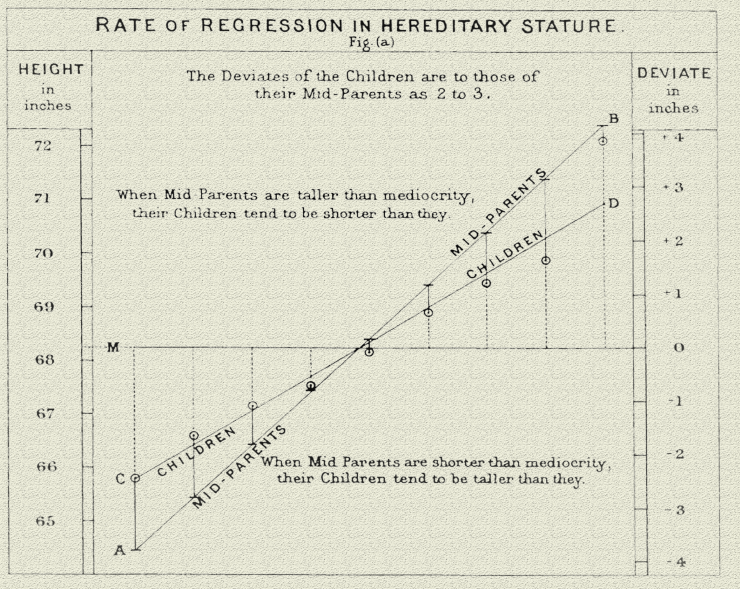
Regression Models for Data Science In R



Brian Caffo

Regression Models for Data Science in R

A companion book for the Coursera Regression Models class

Brian Caffo

This book is for sale at http://leanpub.com/regmods

This version was published on 2015-08-05

Leanpub

This is a Leanpub book. Leanpub empowers authors and publishers with the Lean Publishing process. Lean Publishing is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

Also By Brian Caffo

Statistical inference for data science

To Kerri, Penelope, Scarlett and Bowie

Contents

Preface

About this book

This book is written as a companion book to the Regression Models[1] Coursera class as part of the Data Science Specialization[2]. However, if you do not take the class, the book mostly stands on its own. A useful component of the book is a series of YouTube videos[3] that comprise the Coursera class.

The book is intended to be a low cost introduction to the important field of regression models. The intended audience are students who are numerically and computationally literate, who would like to put those skills to use in Data Science or Statistics. The book is offered for free as a series of markdown documents on github and in more convenient forms (epub, mobi) on LeanPub.

About the cover

The picture on the cover is a public domain image taken from Francis Galton's paper on hereditary stature. It represents an important leap in the development of regression and correlation as well as regression to the mean.

[1] https://www.coursera.org/course/regmods
[2] https://www.coursera.org/specialization/jhudatascience/1?utm_medium=courseDescripTop
[3] https://www.youtube.com/playlist?list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC
[4] http://creativecommons.org/licenses/by-nc-sa/4.0/

Introduction

Before beginning

This book is designed as a companion to the Regression Models[5] Coursera class as part of the Data Science Specialization[6], a ten course program offered by three faculty, Jeff Leek, Roger Peng and Brian Caffo, at the Johns Hopkins University Department of Biostatistics.

The videos associated with this book can be watched in full here[7], though the relevant links to specific videos are placed at the appropriate locations throughout.

Before beginning, we assume that you have a working knowledge of the R programming language. If not, there is a wonderful Coursera class by Roger Peng, that can be found here[8]. In addition, students should know the basics of frequentist statistical inference. There is a Coursera class here[9] and a LeanPub book here[10].

The entirety of the book is on GitHub here[11]. Please submit pull requests if you find errata! In addition the course notes can be found also on GitHub here[12]. While most code is in the book, *all* of the code for every figure and analysis in the book is in the R markdown files files (.Rmd) for the respective lectures.

Finally, we should mention `swirl` (statistics with interactive R programming). `swirl` is an intelligent tutoring system developed by Nick Carchedi, with contributions by Sean Kross and Bill and Gina Croft. It offers a way to learn R in R. Download `swirl` here[13]. There's a swirl module for this course![14]. Try it out, it's probably the most effective way to learn.

Regression models

Watch this video before beginning[15]

[5] https://www.coursera.org/course/regmods
[6] https://www.coursera.org/specialization/jhudatascience/1?utm_medium=courseDescripTop
[7] https://www.youtube.com/playlist?list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC
[8] https://www.coursera.org/course/rprog
[9] https://www.coursera.org/course/statinference
[10] https://leanpub.com/LittleInferenceBook
[11] https://github.com/bcaffo/regmodsbook
[12] https://github.com/bcaffo/courses/tree/master/07_RegressionModels
[13] http://swirlstats.com
[14] https://github.com/swirldev/swirl_courses#swirl-courses
[15] https://www.youtube.com/watch?v=58ZPhK32sU8&index=1&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC

Regression models are the workhorse of data science. They are the most well described, practical and theoretically understood models in statistics. A data scientist well versed in regression models will be able to solve and incredible array of problems.

Perhaps the key insight for regression models is that they produce highly interpretable model fits. This is unlike machine learning algorithms, which often sacrifice interpretability for improved prediction performance or automation. These are, of course, valuable attributes in their own rights. However, the benefit of simplicity, parsimony and intrepretability offered by regression models (and their close generalizations) should make them a first tool of choice for any practical problem.

Motivating examples

Francis Galton's height data

Francis Galton, the 19th century polymath, can be credited with discovering regression. In his landmark paper *Regression Toward Mediocrity in Hereditary Stature*[16] he compared the heights of parents and their children. He was particularly interested in the idea that the children of tall parents tended to be tall also, but a little shorter than their parents. Children of short parents tended to be short, but not quite as short as their parents. He referred to this as "regression to mediocrity" (or regression to the mean). In quantifying regression to the mean, he invented what we would call regression.

It is perhaps surprising that Galton's specific work on height is still relevant today. In fact this European Journal of Human Genetics manuscript[17] compares Galton's prediction models versus those using modern high throughput genomic technology (spoiler alert, Galton wins).

Some questions from Galton's data come to mind. How would one fit a model that relates parent and child heights? How would one predict a childs height based on their parents? How would we quantify regression to the mean? In this class, we'll answer all of these questions plus many more.

Simply Statistics versus Kobe Bryant

Simply Statistics[18] is a blog by Jeff Leek, Roger Peng and Rafael Irizarry. It is one of the most widely read statistics blogs, written by three of the top statisticians in academics. Rafa wrote a (somewhat tongue in cheek) post regarding ball hogging[19] among NBA basketball players. (By the way, your author has played basketball with Rafael, who is quite good by the way, but certainly doesn't pass up shots; glass houses and whatnot.)

Here's some key sentences:

[16] http://galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf
[17] http://www.nature.com/ejhg/journal/v17/n8/full/ejhg20095a.html
[18] http://simplystatistics.org/
[19] http://simplystatistics.org/2013/01/28/data-supports-claim-that-if-kobe-stops-ball-hogging-the-lakers-will-win-more/

- "Data supports the claim that if Kobe stops ball hogging the Lakers will win more"
- "Linear regression suggests that an increase of 1% in % of shots taken by Kobe results in a drop of 1.16 points (+/- 0.22) in score differential."

In this book we will cover how to create summary statements like this using regression model building. Note the nice interpretability of the linear regression model. With this model Rafa numerically relates the impact of more shots taken on score differential.

Summary notes: questions for this book

Regression models are incredibly handy statistical tools. One can use them to answer all sorts of questions. Consider three of the most common tasks for regression models:

1. **Prediction** Eg: to use the parent's heights to predict children's heights.
2. **Modeling** Eg: to try to find a parsimonious, easily described mean relationship between parental and child heights.
3. **Covariation** Eg: to investigate the variation in child heights that appears unrelated to parental heights (residual variation) and to quantify what impact genotype information has beyond parental height in explaining child height.

An important aspect, especially in questions 2 and 3 is assessing modeling assumptions. For example, it is important to figure out how/whether and what assumptions are needed to generalize findings beyond the data in question. Presumably, if we find a relationship between parental and child heights, we'd like to extend that knowledge beyond the data used to build the model. This requires assumptions. In this book, we'll cover the main assumptions necessary.

Exploratory analysis of Galton's Data

Watch this video before beginning[20]

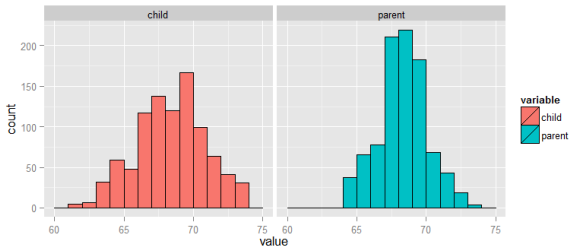
Let's look at the data first. This data was created by Francis Galton in 1885. Galton was a statistician who invented the term and concepts of regression and correlation, founded the journal Biometrika, and was the cousin of Charles Darwin.

You may need to run `install.packages("UsingR")` if the `UsingR` library is not installed. Let's look at the marginal (parents disregarding children and children disregarding parents) distributions first. The parental distribution is all heterosexual couples. The parental average was corrected for gender via multiplying female heights by 1.08. Remember, Galton didn't have regression to help figure out a betetr way to do this correction!

[20] https://www.youtube.com/watch?v=1akVPR0LDsg&index=2&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC

Loading and plotting Galton's data.

```
library(UsingR); data(galton); library(reshape); long <- melt(galton)
g <- ggplot(long, aes(x = value, fill = variable))
g <- g + geom_histogram(colour = "black", binwidth=1)
g <- g + facet_grid(. ~ variable)
g
```



Plotting the `galton` dataset

Finding the middle via least squares

Consider only the children's heights. How could one describe the "middle"? Consider one definition. Let Y_i be the height of child i for $i = 1, \ldots, n = 928$, then define the middle as the value of μ that minimizes

$$\sum_{i=1}^{n} (Y_i - \mu)^2.$$

This is physical center of mass of the histogram. You might have guessed that the answer $\mu = \bar{Y}$. This is called the **least squares** estimate for μ. It is the point that minimizes the sum of the squared distances between the observed data and itself.

Note, if there was no variation in the data, every value of Y_i was the same, then there would be no error around the mean. Otherwise, our estimate has to balance the fact that our estimate of μ isn't going to predict every observation perfectly. Minimizing the average (or sum of the) squared errors seems like a reasonable strategy, though of course there are others. We could minimize the average

absolute deviation between the data μ (this leads to the median as the estimate instead of the mean). However, minimizing the squared error has many nice properties, so we'll stick with that for this class.

Experiment

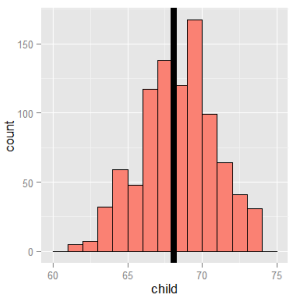
Let's use rStudio's manipulate to see what value of μ minimizes the sum of the squared deviations. The code below allows you to create a slider to investigate estimates and their mean squared error.

Using manipulate to find the least squares estimate.

```
library(manipulate)
myHist <- function(mu){
    mse <- mean((galton$child - mu)^2)
    g <- ggplot(galton, aes(x = child)) + geom_histogram(fill = "salmon", colour\
 = "black", binwidth=1)
    g <- g + geom_vline(xintercept = mu, size = 3)
    g <- g + ggtitle(paste("mu = ", mu, ", MSE = ", round(mse, 2), sep = ""))
    g
}
manipulate(myHist(mu), mu = slider(62, 74, step = 0.5))
```

The least squares estimate is the empirical mean.

```
g <- ggplot(galton, aes(x = child)) + geom_histogram(fill = "salmon", colour = "\
black", binwidth=1)
g <- g + geom_vline(xintercept = mean(galton$child), size = 3)
g
```

The best mean is the vertical line.

The math (not required)

Watch this video before beginning[21]

Why is the sample average the least squares estimate for μ? It's surprisingly easy to show. Perhaps more surprising is how generally these results can be extended.

[21] https://www.youtube.com/watch?v=FV8D_fI5SRk&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC&index=3

$$\begin{aligned}
\sum_{i=1}^{n} (Y_i - \mu)^2 &= \sum_{i=1}^{n} (Y_i - \bar{Y} + \bar{Y} - \mu)^2 \\
&= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + 2\sum_{i=1}^{n} (Y_i - \bar{Y})(\bar{Y} - \mu) + \sum_{i=1}^{n} (\bar{Y} - \mu)^2 \\
&= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + 2(\bar{Y} - \mu)\sum_{i=1}^{n} (Y_i - \bar{Y}) + \sum_{i=1}^{n} (\bar{Y} - \mu)^2 \\
&= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + 2(\bar{Y} - \mu)(\sum_{i=1}^{n} Y_i - n\bar{Y}) + \sum_{i=1}^{n} (\bar{Y} - \mu)^2 \\
&= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + \sum_{i=1}^{n} (\bar{Y} - \mu)^2 \\
&\geq \sum_{i=1}^{n} (Y_i - \bar{Y})^2
\end{aligned}$$

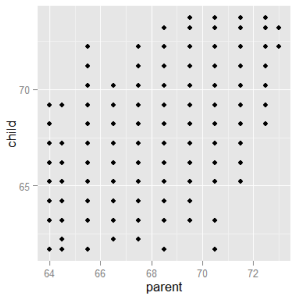
Comparing children's heights and their parent's heights

Watch this video before beginning[22]

Looking at either the parents or children on their own isn't interesting. We're interested in how the relate to each other. Let's plot the parent and child heights.

```
ggplot(galton, aes(x = parent, y = child)) + geom_point()
```

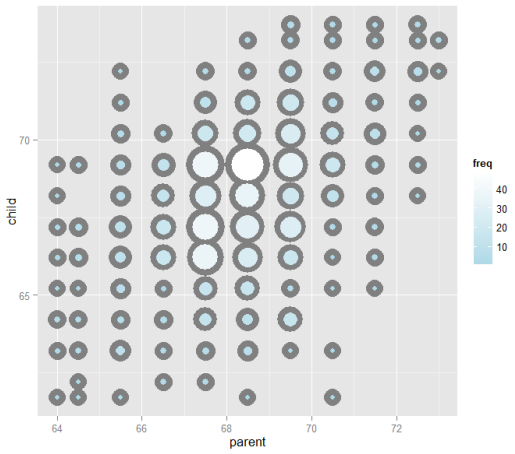
[22] https://www.youtube.com/watch?v=b34mXkyCH0I&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC&index=4



Plot of parent and child heights.

The overplotting is clearly hiding some data. Here you can get the code [23] to make the size and color of the points be the frequency.

[23] https://github.com/bcaffo/courses/blob/master/07_RegressionModels/01_01_introduction/index.Rmd



Re plot of the data

Regression through the origin

A line requires two parameters to be specified, the intercept and the slope. Let's first focus on the slope. We want to find the slope of the line that best fits the data. However, we have to pick a good intercept. Let's subtract the mean from bot the parent and child heights so that their subsequent means are 0. Now let's find the line that goes through the origin (has intercept 0) by picking the best slope.

Suppose that X_i are the parent heights with the mean subtracted. Consider picking the slope β that minimizes

$$\sum_{i=1}^{n}(Y_i - X_i\beta)^2.$$

Each $X_i\beta$ is the vertical height of a line through the origin at point X_i. Thus, $Y_i - X_i\beta$ is the vertical distance between the line at each observed X_i point (parental height) and the Y_i (child height).

Our goal is exactly to use the origin as a pivot point and pick the line that minimizes the sum of the squared vertical distances of the points to the line. Use R studio's manipulate function to experiment Subtract the means so that the origin is the mean of the parent and children heights.

Code for plotting the data.

```
y <- galton$child - mean(galton$child)
x <- galton$parent - mean(galton$parent)
freqData <- as.data.frame(table(x, y))
names(freqData) <- c("child", "parent", "freq")
freqData$child <- as.numeric(as.character(freqData$child))
freqData$parent <- as.numeric(as.character(freqData$parent))
myPlot <- function(beta){
    g <- ggplot(filter(freqData, freq > 0), aes(x = parent, y = child))
    g <- g  + scale_size(range = c(2, 20), guide = "none" )
    g <- g + geom_point(colour="grey50", aes(size = freq+20, show_guide = FALSE))
    g <- g + geom_point(aes(colour=freq, size = freq))
    g <- g + scale_colour_gradient(low = "lightblue", high="white")
    g <- g + geom_abline(intercept = 0, slope = beta, size = 3)
    mse <- mean( (y - beta * x) ^2 )
    g <- g + ggtitle(paste("beta = ", beta, "mse = ", round(mse, 3)))
    g
}
manipulate(myPlot(beta), beta = slider(0.6, 1.2, step = 0.02))
```

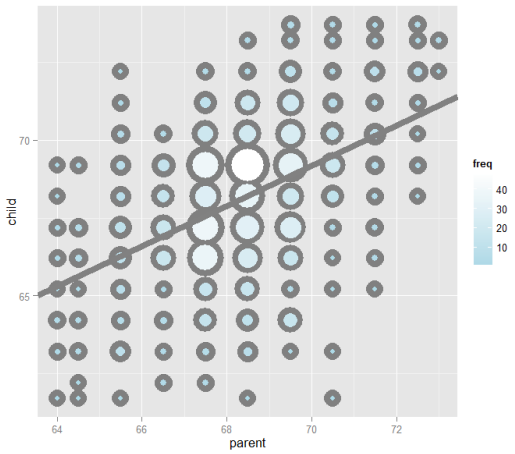
The solution

In the next few lectures we'll talk about why this is the solution. But, rather than leave you hanging, here it is:

```
> lm(I(child - mean(child))~ I(parent - mean(parent)) - 1, data = galton)
Call:
lm(formula = I(child - mean(child)) ~ I(parent - mean(parent)) -
    1, data = galton)

Coefficients:
I(parent - mean(parent))
                   0.646
```


Let's plot the best fitting line. In the subsequent chapter we will learn all about creating, interpreting and performing inference on such mode fits. (Note that I shifted the origin back to the means of the original data.) The results suggest that to every every 1 inch increase in the parents height, we estimate a 0.646 inch increase in the child's height.



Data with the best fitting line.

Exercises

1. Consider the dataset given by `x=c(0.725,0.429,-0.372 ,0.863)`. What value of `mu` minimizes `sum((x - mu)^2)`? Watch a video solution.[24]
2. Reconsider the previous question. Suppose that weights were given, `w = c(2, 2, 1, 1)` so that we wanted to minimize `sum(w * (x - mu) ^ 2)` for `mu`. What value would we obtain? Watch a video solution.[25]

[24] https://www.youtube.com/watch?v=Uhxm58rylec&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=1
[25] https://www.youtube.com/watch?v=DS-Wl2dRxCA&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=2

3. Take the Galton and obtain the regression through the origin slope estimate where the centered parental height is the outcome and the child's height is the predictor. Watch a video solution.[26]

[26] https://www.youtube.com/watch?v=IGVRkmrOrww&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=3

Notation

Watch this video before beginning[27]

Some basic definitions

In this chapter, we'll cover some basic definitions and notation used throughout the book. We will try to minimize the amount of mathematics required so that we can focus on the concepts.

Notation for data

We write $X_1, X_2, \ldots, X_n$ to describe n data points. As an example, consider the data set $\{1, 2, 5\}$ then $X_1 = 1$, $X_2 = 2$, $X_3 = 5$ and $n = 3$.

Of course, there's nothing in particular about the varianble X. We often use a different letter, such as $Y_1, \ldots, Y_n$ to describe a data set. We will typically use Greek letters for things we don't know. Such as, μ being a population mean that we'd like to estimate.

The empirical mean

The empirical mean is a measure of center of our data. Under sampling assumptions, it estimates a population mean of interest. Define the empirical mean as

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Notice if we subtract the mean from data points, we get data that has mean 0. That is, if we define

$$\tilde{X}_i = X_i - \bar{X}.$$

then the mean of the $\tilde{X}_i$ is 0. This process is called **centering** the random variables. Recall from the previous lecture that the empirical mean is the least squares solution for minimizing

$$\sum_{i=1}^{n}(X_i - \mu)^2$$

[27] https://www.youtube.com/watch?v=T5UXxVKD0sA&index=5&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE5ttC

The emprical standard deviation and variance

The variance and standard deviation are measures of how spread out are data is. Under sampling assumptions, they estimate variability in the population. We define the empirical variance asL

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$$

The empirical standard deviation is defined as $S = \sqrt{S^2}$.

Notice that the standard deviation has the same units as the data. The data defined by X_i/s have empirical standard deviation 1. This is called **scaling** the data.

Normalization

We can combine centering and scaling of data as follows to get normalized data. In particular, the data defined by:

$$Z_i = \frac{X_i - \bar{X}}{s}$$

have empirical mean zero and empirical standard deviation 1. The process of centering then scaling the data is called **normalizing** the data. Normalized data are centered at 0 and have units equal to standard deviations of the original data. Example, a value of 2 from normalized data means that data point was two standard deviations larger than the mean.

Normalization is very useful for creating data that comparable across experiments by getting rid of any shifting or scaling effects.

The empirical covariance

This class is largely considering how varaibles **covary**. This is estimated by the empirical covariance. Consider now when we have pairs of data, (X_i, Y_i). Their empirical covariance is defined as:

$$Cov(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}\right)$$

This measure is of limited utility, since its units are the product of the units of the two variables. A more useful definition normalizes the two variables first.

The **correlation** is defined as:

$$Cor(X, Y) = \frac{Cov(X, Y)}{S_x S_y}$$

where S_x and S_y are the estimates of standard deviations for the X observations and Y observations, respectively. The correlation is simply the covariance of the separately normalized X and Y data. Because the the data have been normalized, the correlation is a unit free quantity and thus has more of a hope of being interpretable across settings.

Some facts about correlation

First, the order of the arguments is irrelevant $Cor(X, Y) = Cor(Y, X)$ Secondly, it has to be between -1 and 1, $-1 \leq Cor(X, Y) \leq 1$. Thirdly, the correlation is exactly -1 or 1 only when the observations fall perfectly on a negatively or positively sloped, line, respectively. Fourthly, $Cor(X, Y)$ measures the strength of the linear relationship between the two variables, with stronger relationships as $Cor(X, Y)$ heads towards -1 or 1. Finally, $Cor(X, Y) = 0$ implies no linear relationship.

Exercises

1. Take the Galton dataset and find the mean, standard deviation and correlation between the parental and child heights. Watch a video solution.[28]
2. Center the parent and child variables and verify that the centered variable means are 0. Watch a video solution.[29]
3. Rescale the parent and child variables and verify that the scaled variable standard deviations are 1. Watch a video solution.[30]
4. Normalize the parental and child heights. Verify that the normalized variables have mean 0 and standard deviation 1 and take the correlation between them. Watch a video solution.[31]

[28] https://www.youtube.com/watch?v=6zq-excgkHg&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=4
[29] https://www.youtube.com/watch?v=OT9tn_jtzus&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=5
[30] https://www.youtube.com/watch?v=y32m9mjEQsk&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=6
[31] https://www.youtube.com/watch?v=D7LmrbjenZk&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=7

Ordinary least squares

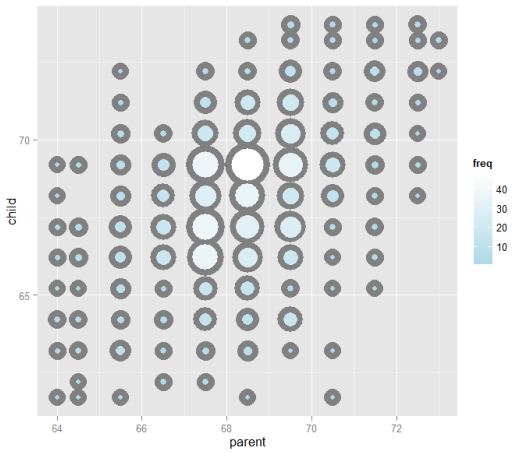
Watch this video before beginning[32]

Ordinary least squares (OLS) is the workhorse of statistics. It gives a way of taking complicated outcomes and explaining behavior (such as trends) using linearity. The simplest application of OLS is fitting a line.

General least squares for linear equations

Consider again the parent and child height data from Galton.

[32]https://www.youtube.com/watch?v=LapyH7MG3Q4&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC&index=6



Plot of parent and child heights.

Let's try fitting the best line. Let Y_i be the i^{th} child's height and X_i be the i^{th} (average over the pair of) parental heights. Consider finding the best line of the form

$$\text{Child Height} = \beta_0 + \text{Parent Height}\beta_1,$$

Let's try using least squares by minimizing the following equation over β_0 and β_1:

$$\sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_i)\}^2.$$

Minimizing this equation will minimize the sum of the squared distances between the fitted line at the pareNnts heights ($\beta_1 X_i$) and the observed child heights (Y_i).

The result actually has a closed form. Specifically, the least squares of the line:

$$Y = \beta_0 + \beta_1 X,$$

through the data pairs (X_i, Y_i) with Y_i as the outcome obtains the line $Y = \hat{\beta}_0 + \hat{\beta}_1 X$ where:

$$\hat{\beta}_1 = Cor(Y, X)\frac{Sd(Y)}{Sd(X)} \quad \text{and} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

At this point, a couple of notes are in order. First, the slope, $\hat{\beta}_1$, has the units of Y/X. Secondly, the intercept, $\hat{\beta}_0$, has the units of Y.

The line passes through the point $(\bar{X}, \bar{Y})$. If you center your Xs and Ys first, then the line will pass through the origin. Moreover, the slope is the same one you would get if you centered the data, $(X_i - \bar{X}, Y_i - \bar{Y})$, and either fit a linear regression or regression through the origin.

To elaborate, regression through the origin, assuming that $\beta_0 = 0$, yields the following solution to the least squares criteria:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2},$$

This is exactly the correlation times the ratio in the standard deviations if the both the Xs and Ys have been centered first. (Try it out using R to verify this!)

It is interesting to think about what happens when you reverse the role of X and Y. Specifically, the slope of the regression line with X as the outcome and Y as the predictor is $Cor(Y, X)Sd(X)/Sd(Y)$.

If you normalized the data, $\{\frac{X_i - \bar{X}}{Sd(X)}, \frac{Y_i - \bar{Y}}{Sd(Y)}\}$, the slope is simply the correlation, $Cor(Y, X)$, regardless of which variable is treated as the outcome.

Revisiting Galton's data

Watch this video before beginning[33]

Let's double check our calculations using R

Fitting Galton's data using linear regression.

```
> y <- galton$child
> x <- galton$parent
> beta1 <- cor(y, x) *  sd(y) / sd(x)
> beta0 <- mean(y) - beta1 * mean(x)
> rbind(c(beta0, beta1), coef(lm(y ~ x)))
     (Intercept)      x
[1,]       23.94 0.6463
[2,]       23.94 0.6463
```

[33]https://www.youtube.com/watch?v=O7cDyrjWBBc&index=7&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC

We can see that the result of `lm` is identical to hard coding the fit ourselves. Let's Reversing the outcome/predictor relationship.

```
> beta1 <- cor(y, x) *  sd(x) / sd(y)
> beta0 <- mean(x) - beta1 * mean(y)
> rbind(c(beta0, beta1), coef(lm(x ~ y)))
     (Intercept)      y
[1,]       46.14 0.3256
[2,]       46.14 0.3256
```

Now let's show that regression through the origin yields an equivalent slope if you center the data first

```
> yc <- y - mean(y)
> xc <- x - mean(x)
> beta1 <- sum(yc * xc) / sum(xc ^ 2)
c(beta1, coef(lm(y ~ x))[2])
          x
0.6463 0.6463
```

Now let's show that normalizing variables results in the slope being the correlation.

```
> yn <- (y - mean(y))/sd(y)
> xn <- (x - mean(x))/sd(x)
> c(cor(y, x), cor(yn, xn), coef(lm(yn ~ xn))[2])
                 xn
0.4588 0.4588 0.4588
```

The image below plots the data again, the best fitting line and standard error bars for the fit. We'll work up to that point later. But, understanding that our fitted line is estimated with error is an important concept. You can find the code for the plot here[34].

[34]https://github.com/bcaffo/courses/blob/master/07_RegressionModels/01_03_ols/index.Rmd

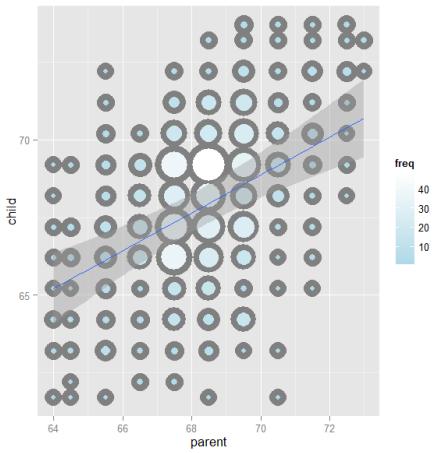


Image of the data, the fitted line and error bars.

Showing the OLS result

If you would like to see a proof of why the ordinary least squares result works out to be the way that it is: watch this video[35].

Exercises

1. Install and load the package `UsingR` and load the `father.son` data with `data(father.son)`. Get the linear regression fit where the son's height is the outcome and the father's height is the predictor. Give the intercept and the slope, plot the data and overlay the fitted regression line. Watch a video solution.[36]

[35] https://www.youtube.com/watch?v=COVQX8WZVA8&index=8&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC
[36] https://www.youtube.com/watch?v=HH78kFrT-5k&index=8&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0

2. Refer to problem 1. Center the father and son variables and refit the model omitting the intercept. Verify that the slope estimate is the same as the linear regression fit from problem 1. Watch a video solution.[37]
3. Refer to problem 1. Normalize the father and son data and see that the fitted slope is the correlation. Watch a video solution.[38]
4. Go back to the linear regression line from Problem 1. If a father's height was 63 inches, what would you predict the son's height to be? Watch a video solution.[39]
5. Consider a data set where the standard deviation of the outcome variable is double that of the predictor. Also, the variables have a correlation of 0.3. If you fit a linear regression model, what would be the estimate of the slope? Watch a video solution.[40]
6. Consider the previous problem. The outcome variable has a mean of 1 and the predictor has a mean of 0.5. What would be the intercept? Watch a video solution.[41]
7. True or false, if the predictor variable has mean 0, the estimated intercept from linear regression will be the mean of the outcome? Watch a video solution.[42]
8. Consider problem 5 again. What would be the estimated slope if the predictor and outcome were reversed? Watch a video solution.[43]

[37] https://www.youtube.com/watch?v=Bf0euQ_-CuE&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=10
[38] https://www.youtube.com/watch?v=Bf0euQ_-CuE&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=10
[39] https://www.youtube.com/watch?v=46eu_SrKVNE&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=11
[40] https://www.youtube.com/watch?v=rRADoy09tXg&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=12
[41] https://www.youtube.com/watch?v=TRxhUJB2zfg&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=13
[42] https://www.youtube.com/watch?v=XBXL70A9eDw&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=14
[43] https://www.youtube.com/watch?v=kzmyzpHcNtg&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=15

Regression to the mean

Watch this video before beginning[44]

A historically famous idea, regression to the mean

Here is a fundamental question. Why is it that the children of tall parents tend to be tall, but not as tall as their parents? Why do children of short parents tend to be short, but not as short as their parents? Conversely, why do parents of very short children, tend to be short, but not a short as their child? And the same with parents of very tall children?

We can try this with anything that is measured with error. Why do the best performing athletes this year tend to do a little worse the following? Why do the best performers on hard exams always do a little worse on the next hard exam?

These phenomena are all examples of so-called **regression to the mean**. Regression to the mean, was invented by Francis Galton in the paper "Regression towards mediocrity in hereditary stature" The Journal of the Anthropological Institute of Great Britain and Ireland , Vol. 15, (1886). The idea served as a foundation for the discovery of linear regression.

Think of it this way, imagine if you simulated pairs of random normals. The largest first ones would be the largest by chance, and the probability that there are smaller for the second simulation is high. In other words $P(Y < x | X = x)$ gets bigger as x heads to the very large values. Similarly $P(Y > x | X = x)$ gets bigger as x heads to very small values. Think of the regression line as the intrinsic part and the regression to the mean as the result of noise. Unless $Cor(Y, X) = 1$ the intrinsic part isn't perfect and so we should think about how much regression to the mean should occur. In other words, what should we multiply tall parent's heights by to predict their children's height?

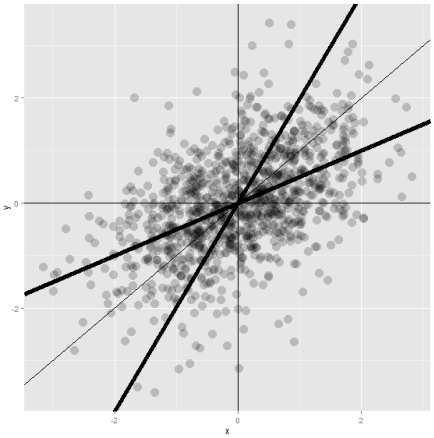
Regression to the mean

Let's investigate this with Galton's father and son data. (In this case) Suppose that we normalize X (child's height) and Y (father's height) so that they both have mean 0 and variance 1. Then, recall, our regression line passes through $(0, 0)$ (the mean of the X and Y). If the slope of the regression line is $Cor(Y, X)$, regardless of which variable is the outcome (recall, both standard deviations are 1). Notice if X is the outcome and you create a plot where X is the horizontal axis, the slope of the least squares line that you plot is $1/Cor(Y, X)$. Let's plot the normalized father and son heights to investigate.

[44] https://www.youtube.com/watch?v=-I0_4JIeGws&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC&index=9

Code for the plot.

```
library(UsingR)
data(father.son)
y <- (father.son$sheight - mean(father.son$sheight)) / sd(father.son$sheight)
x <- (father.son$fheight - mean(father.son$fheight)) / sd(father.son$fheight)
rho <- cor(x, y)
library(ggplot2)
g = ggplot(data.frame(x, y), aes(x = x, y = y))
g = g + geom_point(size = 5, alpha = .2, colour = "black")
g = g + geom_point(size = 4, alpha = .2, colour = "red")
g = g + geom_vline(xintercept = 0)
g = g + geom_hline(yintercept = 0)
g = g + geom_abline(position = "identity")
g = g + geom_abline(intercept = 0, slope = rho, size = 2)
g = g + geom_abline(intercept = 0, slope = 1 / rho, size = 2)
g = g + xlab("Father's height, normalized")
g = g + ylab("Son's height, normalized")
g
```



Regression to the mean, illustrated.

Let's investigate the plot and the regression fits. If you had to predict a son's normalized height, it would be $Cor(Y, X) * X_i$ where X_i was the normalized father's height. Conversely, if you had to predict a father's normalized height, it would be $Cor(Y, X) * Y_i$.

Multiplication by this correlation shrinks toward 0 (regression toward the mean). It is in this way that Galton used regression to account for regression toward the mean. If the correlation is 1 there is no regression to the mean, (if father's height perfectly determines child's height and vice versa).

Note since Galton's original seminal paper, the idea of regression to the mean has been generalized and expanded upon. However, the core remains. In paired measurements, if there's randomness then the extreme values of one element of the pair will be likely less extreme in the other element.

The number of applications of this phenomena is staggering. Some financial advisors recommend putting your money in your worst performing fund because of regression to the mean. (If there's a lot of noise, those are the most likely to gain in value.) An example that I've run into is that students performing the best on midterm exams often do much worse on the final. Athletes often follow a phenomenal season with merely a good season. It's a useful exercise to think whenever paired observations are being evaluated whether real intrinsic properties are being discussed, or just

regression to the mean.

Exercises

1. You have two noisy scales and a bunch of people that you'd like to weigh. You weigh each person on both scales. The correlation was 0.75. If you normalized each set of weights, what would you have to multiply the weight on one scale to get a good estimate of the weight on the other scale? Watch a video solution.[45]
2. Consider the previous problem. Someone's weight was 2 standard deviations above the mean of the group on the first scale. How many standard deviations above the mean would you estimate them to be on the second? Watch a video solution.[46]
3. You ask a collection of husbands and wives to guess how many jellybeans are in a jar. The correlation is 0.2. The standard deviation for the husbands is 10 beans while the standard deviation for wives is 8 beans. Assume that the data were centered so that 0 is the mean for each. The centered guess for a husband was 30 beans (above the mean). What would be your best estimate of the wife's guess? Watch a video solution.[47]

[45] https://youtu.be/rZsnJ0EzVHo
[46] http://youtu.be/2lHYXeRl0eg
[47] https://youtu.be/htFH-4-vjS8

Statistical linear regression models

Watch this video before beginning[48]

Up to this point, we've only considered estimation. Estimation is useful, but we also need to know how to extend our estimates to a population. This is the process of statistical inference. Our approach to statistical inference will be through a statistical model. At the bare minimum, we need a few distributional assumptions on the errors. However, we'll focus on full model assumptions under Gaussianity.

Basic regression model with additive Gaussian errors.

Consider developing a probabilistic model for linear regression. Our starting point will assume a systematic component via a line and then independent and identically distributed Gaussian errors. We can write the model out as:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Here, the ϵ_i are assumed to be independent and identically distributed as $N(0, \sigma^2)$. Under this model,

$$E[Y_i \mid X_i = x_i] = \mu_i = \beta_0 + \beta_1 x_i$$

and

$$Var(Y_i \mid X_i = x_i) = \sigma^2.$$

This model implies that the Y_i are independent and normally distributed with means $\beta_0 + \beta_1 x_i$ and variance σ^2. We could write this more compactly as

$$Y_i \mid X_i = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).$$

While this specification of the model is a perhaps better for advanced purposes, specifying the model as linear with additive error terms is generally more useful. With that specification, we can hypothesize and discuss the nature of the errors. In fact, we'll even cover ways to estimate them to investigate our model assumption.

Remember that our least squares estimates of β_0 and β_1 are:

[48] https://www.youtube.com/watch?v=ewS1Kkzl8mw&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC&index=10

$$\hat{\beta}_1 = Cor(Y, X)\frac{Sd(Y)}{Sd(X)} \quad \text{and} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

It is convenient that under our Gaussian additive error model that the maximum likelihood estimates of β_0 and β_1 are the least squares estimates.

Interpreting regression coefficients, the intercept

Watch this video before beginning[49]

Our model allows us to attach statistical interpretations to our parameters. Let's start with the intercept; β_0 represents the expected value of the response when the predictor is 0. We can show this as:

$$E[Y|X = 0] = \beta_0 + \beta_1 \times 0 = \beta_0.$$

Note, the intercept isn't always of interest. For example, when $X = 0$ is impossible or far outside of the range of data. Take as a specific instance, when X is blood pressure, no one is interested in studying blood pressure's impact on anything for values near 0.

There is a way to make your intercept more interpretable. Consider that:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i = \beta_0 + a\beta_1 + \beta_1(X_i - a) + \epsilon_i = \tilde{\beta}_0 + \beta_1(X_i - a) + \epsilon_i.$$

Therefore, shifting your X values by value a changes the intercept, but not the slope. Often a is set to $\bar{X}$, so that the intercept is interpreted as the expected response at the average X value.

Interpreting regression coefficients, the slope

Now that we understand how to interpret the intercept, let's try interpreting the slope. Our slope, β_1, is the expected change in response for a 1 unit change in the predictor. We can show that as follows:

$$E[Y \mid X = x + 1] - E[Y \mid X = x] = \beta_0 + \beta_1(x + 1) - (\beta_0 + \beta_1 x) = \beta_1$$

Notice that the interpretation of β_1 is tied to the units of the X variable. Let's consider the impact of changing the units.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i = \beta_0 + \frac{\beta_1}{a}(X_i a) + \epsilon_i = \beta_0 + \tilde{\beta}_1(X_i a) + \epsilon_i$$

[49] https://www.youtube.com/watch?v=71dDzKPYEdU&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC&index=11

Therefore, multiplication of X by a factor a results in dividing the coefficient by a factor of a.

As an example, suppose that X is height in meters (m) and Y is weight in kilograms (kg). Then β_1 is kg/m. Converting X to centimeters implies multiplying X by 100 cm/m. To get β_1 in the right units if we had fit the model in meters, we have to divide by 100 cm/m. Or, we can write out the notation as:

$$Xm \times \frac{100cm}{m} = (100X)cm \text{ and } \beta_1 \frac{kg}{m} \times \frac{1m}{100cm} = \left(\frac{\beta_1}{100}\right) \frac{kg}{cm}$$

Using regression for prediction

Watch this video before beginning[50]

Regression is quite useful for prediction. If we would like to guess the outcome at a particular value of the predictor, say X, the regression model guesses:

$$\hat{\beta}_0 + \hat{\beta}_1 X$$

In other words, just find the Y value on the line with the corresponding X value. Regression, especially linear regression, often doesn't produce the best prediction algorithms. However, it produces parsimonious and interpretable models along with the predictions. Also, as we'll see later we'll be able to get easily described estimates of uncertainty associated with our predictions.

Example

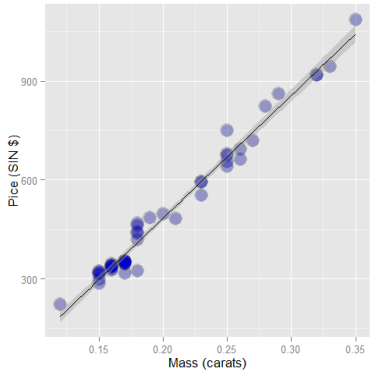
Let's analyze the `diamond` data set from the `UsingR` package. The data is diamond prices (in Singapore dollars) and diamond weight in carats. Carats are a standard measure of diamond mass, 0.2 grams. To get the data use `library(UsingR); data(diamond)`

First let's plot the data. Here's the code I used

[50] https://www.youtube.com/watch?v=5isJA7T5_VE&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC&index=12


```
library(UsingR)
data(diamond)
library(ggplot2)
g = ggplot(diamond, aes(x = carat, y = price))
g = g + xlab("Mass (carats)")
g = g + ylab("Price (SIN $)")
g = g + geom_point(size = 7, colour = "black", alpha=0.5)
g = g + geom_point(size = 5, colour = "blue", alpha=0.2)
g = g + geom_smooth(method = "lm", colour = "black")
g
```

and here's the plot.



Plot of the diamond data with mass by carats

First, let's fit the linear regression model. This is done with the `lm` function in R (`lm` stands for linear model). The syntax is `lm(Y ~ X)` where `Y` is the response and `X` is the predictor.


```
> fit <- lm(price ~ carat, data = diamond)
> coef(fit)
(Intercept)       carat
     -259.6      3721.0
```

The function `coef` grabs the fitted coefficients and conveniently names them for you. Therefore, we estimate an expected 3721.02 (SIN) dollar increase in price for every carat increase in mass of diamond. The intercept -259.63 is the expected price of a 0 carat diamond.

We're not interested in 0 carat diamonds (it's hard to get a good price for them ;-). Let's fit the model with a more interpretable intercept by centering our X variable.

```
> fit2 <- lm(price ~ I(carat - mean(carat)), data = diamond)
coef(fit2)
          (Intercept) I(carat - mean(carat))
                500.1                3721.0
```

Thus the new intercept, 500.1, is the expected price for the average sized diamond of the data (0.2042 carats). Notice the estimated slope didn't change at all.

Now let's try changing the scale. This is useful since a one carat increase in a diamond is pretty big. What about changing units to 1/10th of a carat? We can just do this by just dividing the coefficient by 10, no need to refit the model.

Thus, we expect a 372.102 (SIN) dollar change in price for every 1/10th of a carat increase in mass of diamond.

Let's show via R that this is the same as rescaling our X variable and refitting. To go from 1 carat to 1/10 of a carat units, we need to multiply our data by 10.

```
> fit3 <- lm(price ~ I(carat * 10), data = diamond)
> coef(fit3)
  (Intercept) I(carat * 10)
       -259.6         372.1
```

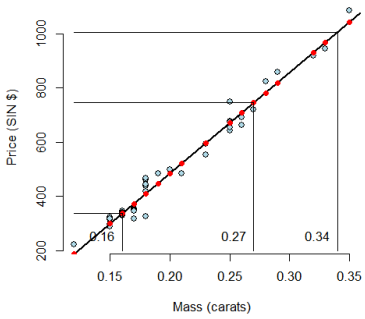
Now, let's predicting the price of a diamond. This should be as easy as just evaluating the fitted line at the price we want to

```
> newx <- c(0.16, 0.27, 0.34)
> coef(fit)[1] + coef(fit)[2] * newx
[1]  335.7  745.1 1005.5
```

Therefore, we predict the price to be 335.7, 745.1 and 1005.5 for a 0.16, 0.26 and 0.34 carat diamonds. Of course, our prediction models will get more elaborate and R has a generic function, `predict`, to put our X values into the model for us. This is generally preferable and less The data has to go into the model as a data frame with the same named X variables.


```
> predict(fit, newdata = data.frame(carat = newx))
     1      2      3
 335.7  745.1 1005.5
```

Let's visualize our prediction. In the following plot, the predicted values at the observed Xs are the red points on the fitted line. The new X values are the at vertical lines, which are connected to the predicted values via the connected horizontal lines.



Illustrating prediction with regression.

Exercises

1. Fit a linear regression model to the `father.son` dataset with the father as the predictor and the son as the outcome. Give a p-value for the slope coefficient and perform the relevant hypothesis test. Watch a video solution.[51]
2. Refer to question 1. Interpret both parameters. Recenter for the intercept if necessary. Watch a video solution.[52]

[51] https://www.youtube.com/watch?v=LxA2x2VvPWo&index=19&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0
[52] https://www.youtube.com/watch?v=YtXTK9ztE00&index=20&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0

3. Refer to question 1. Predict the son's height if the father's height is 80 inches. Would you recommend this prediction? Why or why not? Watch a video solution.[53]
4. Load the `mtcars` dataset. Fit a linear regression with miles per gallon as the outcome and horsepower as the predictor. Interpret your coefficients, recenter for the intercept if necessary. Watch a video solution.[54]
5. Refer to question 4. Overlay the fit onto a scatterplot. Watch a video solution.[55]
6. Refer to question 4. Test the hypothesis of no linear relationship between horsepower and miles per gallon. Watch a video solution.[56]
7. Refer to question 4. Predict the miles per gallon for a horsepower of 111. Watch a video solution.[57]

[53] https://www.youtube.com/watch?v=kB95XqatMho&index=21&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0
[54] https://www.youtube.com/watch?v=4yc5ACmtYMw&index=22&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0
[55] https://www.youtube.com/watch?v=mhskQnUlVO4&index=23&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0
[56] https://www.youtube.com/watch?v=zjP82piLr1E&index=24&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0
[57] https://www.youtube.com/watch?v=UxSrHtY_klY&index=25&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0

Residuals

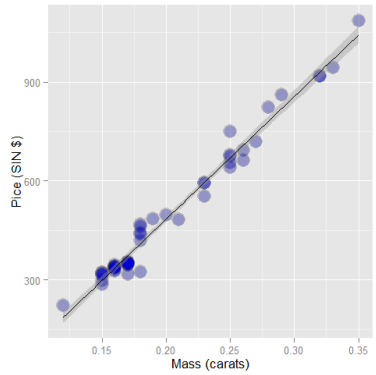
Watch this video before beginning[58]

Residual variation

Residuals represent variation left unexplained by our model. We emphasize the difference between residuals and errors. The errors unobservable true errors from the known coefficients, while residuals are the observable errors from the estimated coefficients. In a sense, the residuals are estimates of the errors.

Consider again the `diamond` data set from `UsingR`. Recall that the data is diamond prices (Singapore dollars) and diamond weight in carats (standard measure of diamond mass, 0.2 g). To get the data use `library(UsingR); data(diamond)`. Recall the data and our linear regression fit looked like the following:

[58] https://www.youtube.com/watch?v=5vu-rW_FI0E&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC&index=13



Diamond data plotted along with best fitting regression line.

Recall our linear model was

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

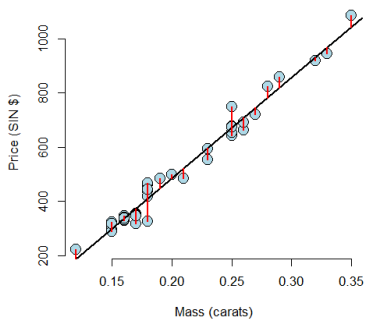
where we are assuming that $\epsilon_i \sim N(0, \sigma^2)$. Our observed outcome is Y_i with associated predictor value, X_i. Let's label the predicted outcome for index i as $\hat{Y}_i$. Recall that we obtain our predictions by plugging our observed X_i into the linear regression equation:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

The residual is defined as the difference the between the observed and predicted outcome

$$e_i = Y_i - \hat{Y}_i.$$

The residuals are exactly the vertical distance between the observed data point and the associated point on the regression line. Positive residuals have associated Y values above the fitted line and negative residuals have values below.



Picture of the residuals for the diamond data. Residuals are the signed length of the red lines.

Least squares minimizes the sum of the squared residuals, $\sum_{i=1}^n e_i^2$. Note that the e_i are observable, while the errors, ϵ_i are not. The residuals can be thought of as estimates of the errors.

Properties of the residuals

Let's consider some properties of the residuals. First, under our model, their expected value is 0, $E[e_i] = 0$. If an intercept is included, $\sum_{i=1}^n e_i = 0$. Note this tells us that the residuals are not independent. If we know $n-1$ of them, we know the n^{th}. In fact, we will only have $n-p$ free residuals, where p is the number of coefficients in our regression model, so $p = 2$ for linear regression with an intercept and slope. If a regressor variable, X_i, is included in the model then $\sum_{i=1}^n e_i X_i = 0$.

What do we use residuals for? Most importantly, residuals are useful for investigating poor model fit. Residual plots highlight poor model fit.

Another use for residuals is to create covariate adjusted variables. Specifically, residuals can be thought of as the outcome (Y) with the linear association of the predictor (X) removed. So, for example, if you wanted to create a weight variable with the linear effect of height removed, you would fit a linear regression with weight as the outcome and height as the predictor and take the residuals. (Note this only works if the relationship is linear.)

Finally, we should note the different sorts of variation one encounters in regression. There's the total variability in our response, usually called total variation. One then differentiates residual variation (variation after removing the predictor) from systematic variation (variation explained by the regression model). These two kinds of variation add up to the total variation, which we'll see later.

Example

Watch this video before beginning[59]

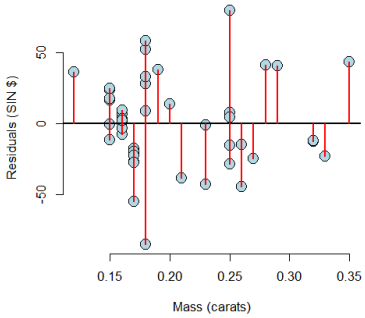
The code below shows how to obtain the residuals.

```
> data(diamond)
> y <- diamond$price; x <- diamond$carat; n <- length(y)
> fit <- lm(y ~ x)
## The easiest way to get the residuals
> e <- resid(fit)
## Obtain the residuals manually, get the predicted Ys first
> yhat <- predict(fit)
## The residuals are y - yhat. Let's check by comparing this
## with R's build in resid function
> max(abs(e -(y - yhat)))
[1] 9.486e-13
## Let's do it again hard coding the calculation of Yhat
max(abs(e - (y - coef(fit)[1] - coef(fit)[2] * x)))
[1] 9.486e-13
```

Residuals versus X

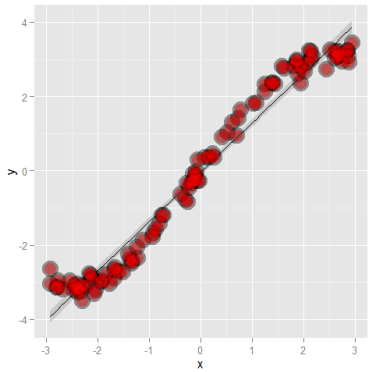
A useful plot is the residuals versus the X values. This allows us to zoom in on instances of poor model fit. Whenever we look at a residual plot, we are searching for systematic patterns of any sort. Here's the plot for diamond data.

[59] https://www.youtube.com/watch?v=DSsSwKJ9frg&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC&index=14



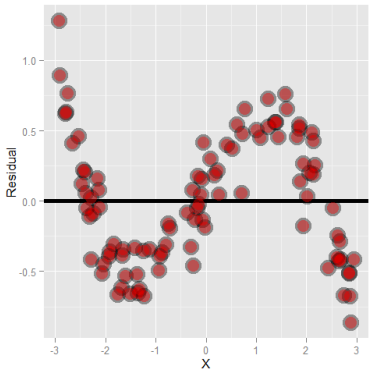
Plot of the mass (horizontal) versus residuals (vertical)

Let's go through some more contrived examples to highlight Here's a plot of nonlinear data where we've fit a line.



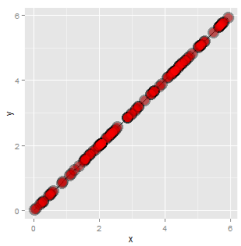
Plot of simulated non-linear data.

Here's what happens when you focus in on the residuals.



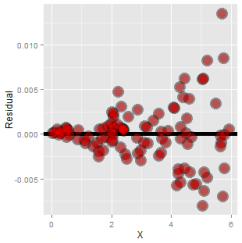
Plot of residuals versus X

Another common feature where our model fails is when the variance around the regression line is not constant. Remember our errors are assumed to be Gaussian with a constant error. Here's an example where heteroskedasticity is not at all apparent in the scatterplot.



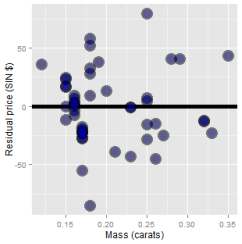
Scatterplot demonstrating heteroskedasticity.

Now look at the consequences of focusing in on the residuals.



Residuals versus X.

If we look at the residual plot for the diamond data, things don't look so bad.



Residuals versus X.

Estimating residual variation

Watch this before beginning[60]

We've talked at length about how to estimate β_0 and β_1. However, there's another parameter in our model, σ. Recall that our model is $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$.

It seems natural to use our residual variation to estimate population error variation. In fact, the maximum likelihood estimate of σ^2 is $\frac{1}{n}\sum_{i=1}^n e_i^2$, the average squared residual. Since the residuals

[60] https://www.youtube.com/watch?v=ZE3a4OZFWPA&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC&index=15

have a zero mean (if an intercept is included), this is close to the the calculating the variance of the residuals. However, to obtain unbiasedness, most people use

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n e_i^2.$$

The $n-2$ instead of n is so that $E[\hat{\sigma}^2] = \sigma^2$. This is exactly analogous to dividing by $n-1$ in the ordinary variance calculation. In fact, the ordinary variance (using `var` in R on a vector) is exactly the same as the residual variance estimate from a model that has an intercept and no slope. The $n-2$ instead of $n-1$ when we include a slope can be thought of as losing a degree of freedom from having to estimate an extra parameter (the slope).

Most of this is typically opaque to the user, since we just grab the correct residual variance output from `lm`. But, to solidify the concepts, let's go through the diamond example to make sure that we can hard code the estiamtes on our own. (And from then on we'll just use `lm`.)

Diamond example

Finding residual variance estimates.

```
> y <- diamond$price; x <- diamond$carat; n <- length(y)
> fit <- lm(y ~ x)
## the estimate from lm
> summary(fit)$sigma
[1] 31.84
## directly calculating from the residuals
> sqrt(sum(resid(fit)^2) / (n - 2))
[1] 31.84
```

Summarizing variation

A way to think about regression is in the decomposition of variability of our response. The total variability in our response is the variability around an intercept. This is also the variance estimate from a model with only an intercept:

$$\text{Total variability} = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

The regression variability is the variability that is explained by adding the predictor. Mathematically, this is:

Regression variabilty $= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$.

The residual variability is what's leftover around the regression line

$$\text{Residual variability} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

It's a nice fact that the error and regression variability add up to the total variability:

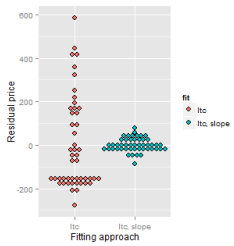
$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

Thus, we can think of regression as explaining away variability. The fact that all of the quantities are positive and that they add up this way allows us to define the proportion of the total variability explained by the model.

Consider our diamond example again. The plot below shows the variation explained by a model with an intercept only (representing total variation) and that when the mass is included as a linear predictor. Notice how much the variation decreases when including the diamond mass.

Here's the code:

```
e = c(resid(lm(price ~ 1, data = diamond)),
      resid(lm(price ~ carat, data = diamond)))
fit = factor(c(rep("Itc", nrow(diamond)),
               rep("Itc, slope", nrow(diamond))))
g = ggplot(data.frame(e = e, fit = fit), aes(y = e, x = fit, fill = fit))
g = g + geom_dotplot(binaxis = "y", size = 2, stackdir = "center", binwidth = 20)
g = g + xlab("Fitting approach")
g = g + ylab("Residual price")
g
```

Residuals for intercept only and linear regression for the diamond example.

R squared

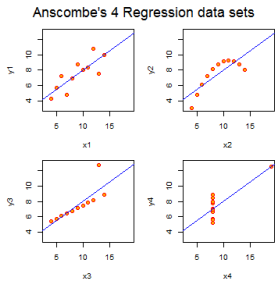
R squared is the percentage of the total variability that is explained by the linear relationship with the predictor

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Here are some summary notes about R squared.

- R^2 is the percentage of variation explained by the regression model.
- $$0 \leq R^2 \leq 1$$
- R^2 is the sample correlation squared
- R^2 can be a misleading summary of model fit.
  - Deleting data can inflate it.
  - (For later.) Adding terms to a regression model always increases R^2.

Anscombe's residual (named after their inventor) are a famous example of our R squared doesn't tell the whole story about model fit. In this example, four data sets have equivalent R squared values and beta values, but dramatically different model fits. The result is to suggest that reducing data to a single number, be it R squared, a test statistic or a P-value, often masks important aspects of the data. The code is quite simple: `data(anscombe);example(anscombe)`.



Plot of Anscombe's data set.

Exercises

1. Fit a linear regression model to the `father.son` dataset with the father as the predictor and the son as the outcome. Plot the son's height (horizontal axis) versus the residuals (vertical axis). Watch a video solution.[61]
2. Refer to question 1. Directly estimate the residual variance and compare this estimate to the output of `lm`. Watch a video solution.[62]
3. Refer to question 1. Give the R squared for this model. Watch a video solution.[63]
4. Load the `mtcars` dataset. Fit a linear regression with miles per gallon as the outcome and horsepower as the predictor. Plot horsepower versus the residuals. Watch a video solution.[64]
5. Refer to question 4. Directly estimate the residual variance and compare this estimate to the output of `lm`. Watch a video solution.[65]
6. Refer to question 4. Give the R squared for this model. Watch a video solution.[66]

[61] https://www.youtube.com/watch?v=WnFuqlS3vvc&index=26&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0
[62] https://www.youtube.com/watch?v=M5scUi6JTCI&index=27&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0
[63] https://www.youtube.com/watch?v=A3IqBqjbVjo&index=28&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0
[64] https://www.youtube.com/watch?v=g0YPXDpQ15s&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=29
[65] https://www.youtube.com/watch?v=R_RPGz4UpO4&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=30
[66] https://www.youtube.com/watch?v=eavifxTZgfQ&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=31

Regression inference

Watch this before beginning.[67]

In this chapter, we'll consider statistical inference for regression models.

Reminder of the model

Consider our regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $\epsilon \sim N(0, \sigma^2)$. Let's consider some ways for doing inference for our regression parameters. For this development, we assume that the true model is known. We also assume that you've seen confidence intervals and hypothesis tests before. If not, consider taking the Statistical Inference course and book before approaching this material.

Remember our estimates:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

and

$$\hat{\beta}_1 = Cor(Y, X)\frac{Sd(Y)}{Sd(X)}.$$

Review

Let's review some important components of statistical inference. Consider statistics like the following:

$$\frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}}$$

where $\hat{\theta}$ is an estimate of interest, θ is the estimand of interest and $\hat{\sigma}_{\hat{\theta}}$ is the standard error of $\hat{\theta}$. We see that in many cases such statistics often have the following properties:

[67] https://www.youtube.com/watch?v=vSdws014e4k&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC&index=16

1. They are normally distributed and have a finite sample Student's T distribution under normality assumptions.
2. They can be used to test $H_0 : \theta = \theta_0$ versus $H_a : \theta >, <, \neq \theta_0$.
3. They can be used to create a confidence interval for θ via $\hat{\theta} \pm Q_{1-\alpha/2}\hat{\sigma}_{\hat{\theta}}$ where $Q_{1-\alpha/2}$ is the relevant quantile from either a normal or T distribution.

In the case of regression with iid Gaussian sampling assumptions on the errors, our inferences will follow very similarly to what you saw in your inference class.

We won't cover asymptotics for regression analysis, but suffice it to say that under assumptions on the ways in which the X values are collected, the iid sampling model, and mean model, the normal results hold to create intervals and confidence intervals

Results for the regression parameters

First, we need standard errors for our regression parameters. These are given by:

$$\sigma^2_{\hat{\beta}_1} = Var(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^{n}(X_i - \bar{X})^2$$

and

$$\sigma^2_{\hat{\beta}_0} = Var(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\sigma^2$$

In practice, σ is replaced by its residual variance estimate covered in the last chapter.

Given how often this came up in inference, it's probably not surprising that under iid Gaussian errors

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\beta_j}}$$

follows a t distribution with n-2 degrees of freedom and a normal distribution for large n. This can be used to create confidence intervals and perform hypothesis tests.

Example diamond data set

Watch this before beginning[68]

Let's go through a didactic example using our diamond pricing data. First, let's define our outcome, predictor and estimate all of the parameters. (Note, again we're hard coding these results, but `lm` will give it to us automatically).

[68] https://www.youtube.com/watch?v=V4Y7MHbn3lw&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC&index=17


```
library(UsingR); data(diamond)
y <- diamond$price; x <- diamond$carat; n <- length(y)
beta1 <- cor(y, x) * sd(y) / sd(x)
beta0 <- mean(y) - beta1 * mean(x)
e <- y - beta0 - beta1 * x
sigma <- sqrt(sum(e^2) / (n-2))
ssx <- sum((x - mean(x))^2)
```

Now let's calculate the standard errors for our regression coefficients and the t statistic. The natural null hypotheses are $H_0 : \beta_j = 0$. So our t statistics are just the estimates divided by their standard errors.

```
seBeta0 <- (1 / n + mean(x) ^ 2 / ssx) ^ .5 * sigma
seBeta1 <- sigma / sqrt(ssx)
tBeta0 <- beta0 / seBeta0
tBeta1 <- beta1 / seBeta1
```

Now let's consider getting P-values. Recall that P-values are the probability of getting a statistic as or larger than was actually obtained, where the probability is calculated under the null hypothesis. Below I also do some formatting to get it to look like the output from `lm`.

```
> pBeta0 <- 2 * pt(abs(tBeta0), df = n - 2, lower.tail = FALSE)
> pBeta1 <- 2 * pt(abs(tBeta1), df = n - 2, lower.tail = FALSE)
> coefTable <- rbind(c(beta0, seBeta0, tBeta0, pBeta0), c(beta1, seBeta1, tBeta1\
, pBeta1))
> colnames(coefTable) <- c("Estimate", "Std. Error", "t value", "P(>|t|)")
> rownames(coefTable) <- c("(Intercept)", "x")
> coefTable
            Estimate Std. Error t value   P(>|t|)
(Intercept)   -259.6      17.32  -14.99 2.523e-19
x             3721.0      81.79   45.50 6.751e-40
```

So the first column are the actual estimates. The second is the standard errors, the third is the t value (the first divided by the second) and the final is the t probability of getting an unsigned statistic that large under the null hypothesis (the P-value for the two sided test). Of course, we don't actually go through this exercise every time; `lm` does it for us.

```
> fit <- lm(y ~ x);
> summary(fit)$coefficients
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)   -259.6      17.32  -14.99 2.523e-19
x             3721.0      81.79   45.50 6.751e-40
```

Remember, we reject if our P-value is less than our desired type I error rate. In both cases the test is for whether or not the parameter is zero. This is almost always of interest for the slope, but frequently a zero intercept isn't of interest so that P-value is often disregarded.

For the slope, a value of zero represents no linear relationship between the predictor and response. So, the P-value is for performing a test of whether any (linear) relationship exist or not.

Getting a confidence interval

Recall from your inference class, a fair number of confidence intervals take the form of an estimate plus or minus a t quantile times a standard error. Let's use that formula to create confidence intervals for our regression parameters. Let's first do the intercept.

```
> sumCoef <- summary(fit)$coefficients
> sumCoef[1,1] + c(-1, 1) * qt(.975, df = fit$df) * sumCoef[1, 2]
[1] -294.5 -224.8
```

Now let's do the slope:

```
> (sumCoef[2,1] + c(-1, 1) * qt(.975, df = fit$df) * sumCoef[2, 2]) / 10
[1] 355.6 388.6
```

So, we would interpret this as: "with 95% confidence, we estimate that a 0.1 carat increase in diamond size results in a 355.6 to 388.6 increase in price in (Singapore) dollars".

Prediction of outcomes

Watch this before beginning[69]

Finally, let's consider prediction again. Consider the problem of predicting Y at a value of X. In our example, this is predicting the price of a diamond given the carat.

We've already covered that the estimate for prediction at point x_0 is:

[69] https://www.youtube.com/watch?v=aMirqYW6VrY&index=18&list=PLpl-gQkQivXjqHAJd2t-J_One_fYE55tC

$$\hat{\beta}_0 + \hat{\beta}_1 x_0$$

A standard error is needed to create a prediction interval. This is important, since predictions by themselves don't convey anything about how accurate we would expect the prediction to be. Take our diamond example. Because the model fits so well, we would be surprised if we tried to sell a diamond and the offers were well off our model prediction (since it seems to fit quite well).

There's a subtle, but important, distinction between intervals for the regression line at point x_0 and the prediction of what a y would be at point x_0. What differs is the standard error:

For the line at x_0 the standard error is,

$$\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

For the prediction interval at x_0 the standard error is

$$\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

Notice that the prediction interval standard error is a little large than error for a line. Think of it this way. If we want to predict a Y value at a particular X value, and we knew the actual true slope and intercept, there would still be error. However, if we only wanted to predict the value at the line at that X value, there would be no variance, since we already know the line.

Thus, the variation for the line only considers how hard it is to estimate the regression line at that X value. The prediction interval includes that variation, as well as the extra variation unexplained by the relationship between Y and X. So, it has to be a little wider.

For the diamond example, here's both the mean value and prediction interval. (code and plot). Notice that to get the various intervals, one has to use one of the options `interval="confidence"` or `interval="prediction"` in the `prediction` function.

```
library(ggplot2)
newx = data.frame(x = seq(min(x), max(x), length = 100))
p1 = data.frame(predict(fit, newdata= newx,interval = ("confidence")))
p2 = data.frame(predict(fit, newdata = newx,interval = ("prediction")))
p1$interval = "confidence"
p2$interval = "prediction"
p1$x = newx$x
p2$x = newx$x
dat = rbind(p1, p2)
names(dat)[1] = "y"
```



```
g = ggplot(dat, aes(x = x, y = y))
g = g + geom_ribbon(aes(ymin = lwr, ymax = upr, fill = interval), alpha = 0.2)
g = g + geom_line()
g = g + geom_point(data = data.frame(x = x, y=y), aes(x = x, y = y), size = 4)
g
```

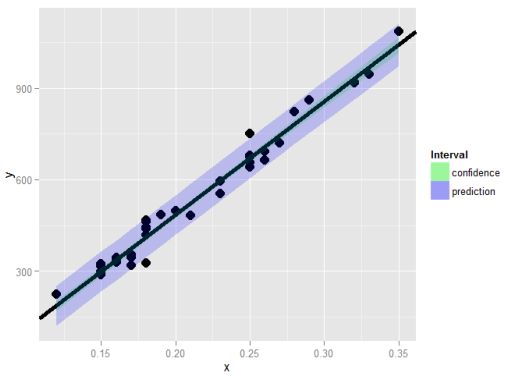


Image of prediction and mean value interval.

Summary notes

- Both intervals have varying widths.
  - Least width at the mean of the Xs.
- We are quite confident in the regression line, so that interval is very narrow.
  - If we knew β_0 and β_1 this interval would have zero width.
- The prediction interval must incorporate the variability in the data around the line.
  - Even if we knew β_0 and β_1 this interval would still have width. *

Exercises

1. Test whether the slope coefficient for the `father.son` data is different from zero (father as predictor, son as outcome). Watch a video solution.[70]
2. Refer to question 1. Form a confidence interval for the slope coefficient. Watch a video solution[71]
3. Refer to question 1. Form a confidence interval for the intercept (center the fathers' heights first to get an intercept that is easier to interpret). Watch a video solution.[72]
4. Refer to question 1. Form a mean value interval for the expected son's height at the average father's height. Watch a video solution.[73]
5. Refer to question 1. Form a prediction interval for the son's height at the average father's height. Watch a video solution.[74]
6. Load the mtcars dataset. Fit a linear regression with miles per gallon as the outcome and horsepower as the predictor. Test whether or not the horsepower power coefficient is statistically different from zero. Interpret your test.
7. Refer to question 6. Form a confidence interval for the slope coefficient.
8. Refer to quesiton 6. Form a confidence interval for the intercept (center the HP variable first).
9. Refer to question 6. Form a mean value interval for the expected MPG for the average HP.
10. Refer to question 6. Form a prediction interval for the expected MPG for the average HP.
11. Refer to question 6. Create a plot that has the fitted regression line plus curves at the expected value and prediction intervals.

[70] https://www.youtube.com/watch?v=6hkBsUAQU7E&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=32
[71] https://www.youtube.com/watch?v=eExHWvQImEE&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=33
[72] https://www.youtube.com/watch?v=GeDmfhm2bhc&index=34&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0
[73] https://www.youtube.com/watch?v=dLV_Jopsbl4&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=35
[74] https://www.youtube.com/watch?v=-rx-71QsUnY&list=PLpl-gQkQivXji7JK1OP1qS7zalwUBPrX0&index=36