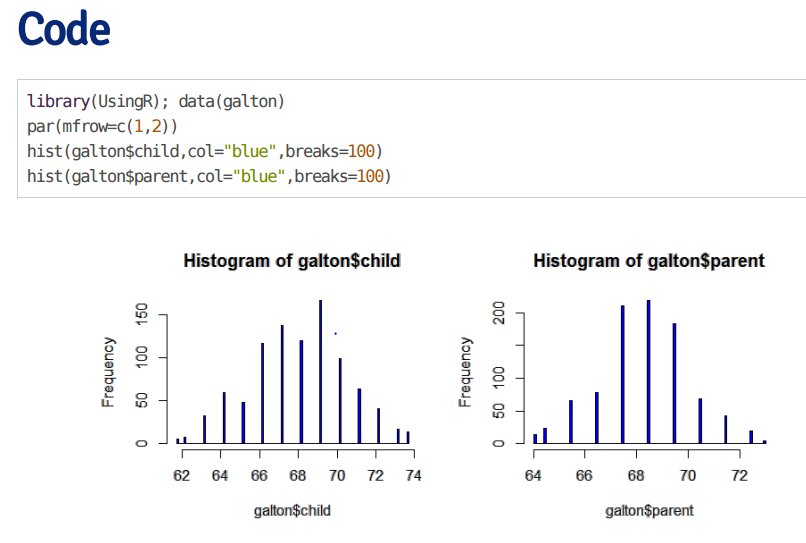
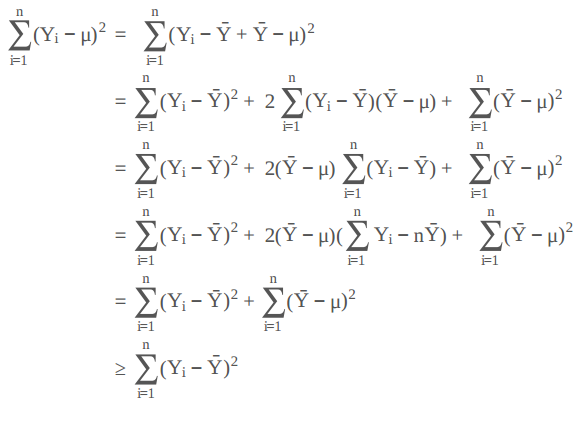
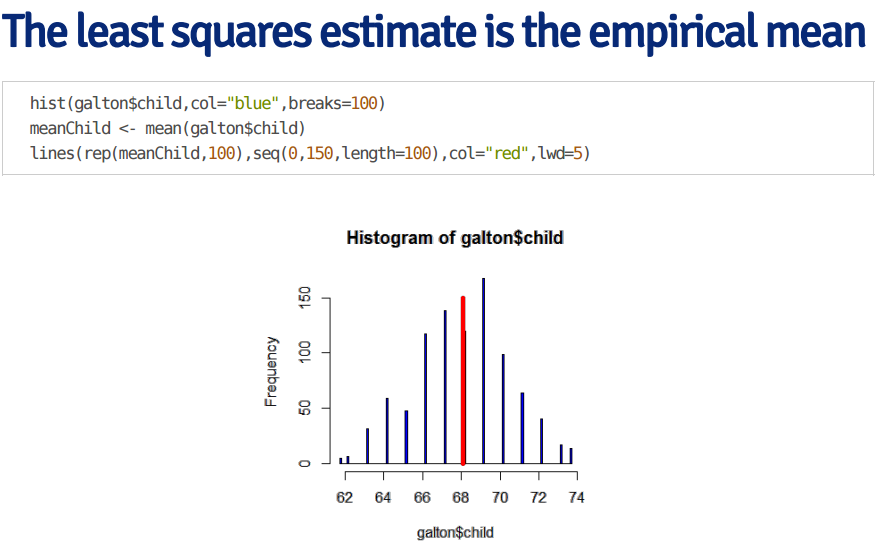
**Coursera Data Science Course 7: Regression Models**

**1. Some Potential Questions for this class:**-to use the parents’ heights to predict childrens’ heights.  
-to try to find a parsimonious, easily described mean relationship between parent and children’s heights.  
-to investigate the variation in childrens’ heights that appears unrelated to parents’ heights (residual variation)  
-to quantify what impact genotype information has beyond parental height in explaining child height  
-to figure out how/whether and what assumptions are needed to generalize findings beyond the data in question  
-why do children of very tall parents tend to be tall, but a little shorter than their parents and why children of very short parents tend to be short, but a little taller than their parents? (regression to the mean)

**2. Galton’s Data**-install.packages(“UsingR”)  
-Ex: Marginal distribution:  
 -Parent distribution is all heterosexual couples  
 -Correction for gender via multiplying female heights by 1.08  
 -Overplotting is an issue from discretization



Consider only the children’s heights.  
-How could one describe the “middle”  
-define the middle as the value of the mean that minimizes the variation  
-you might have guess the answer mean = X

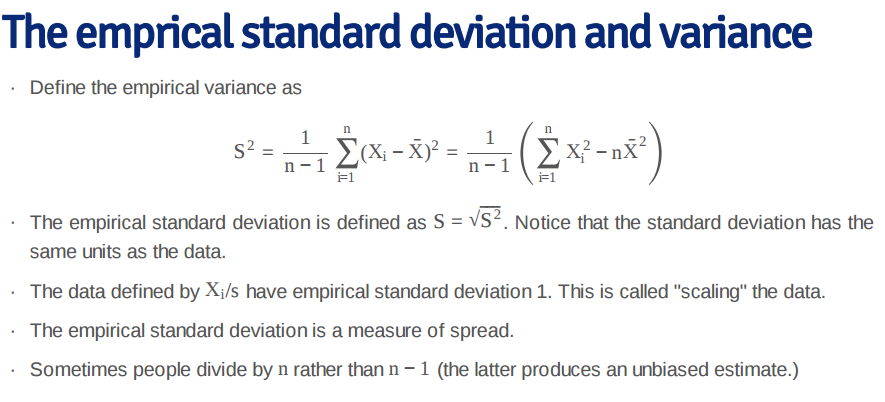
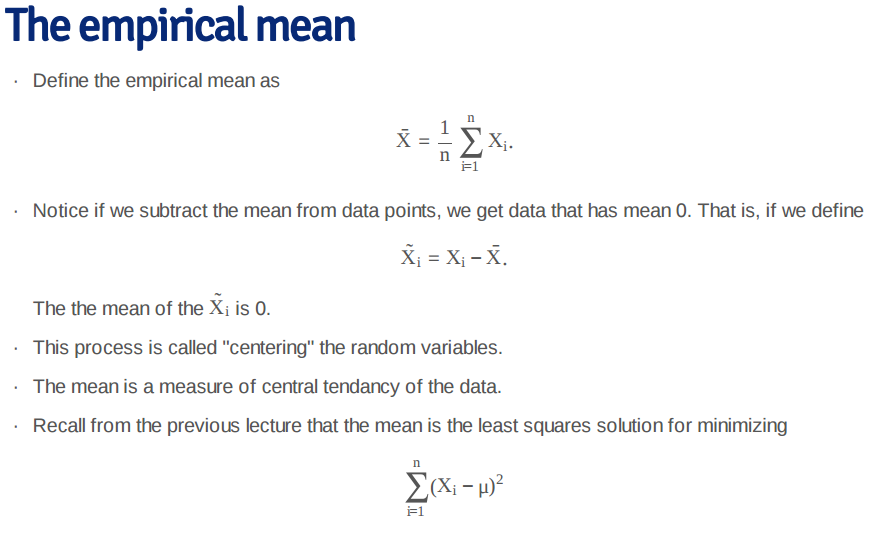
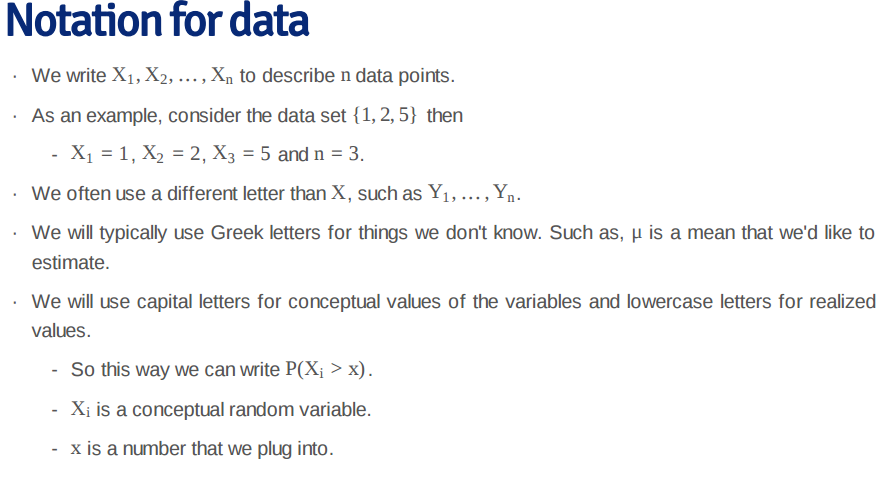


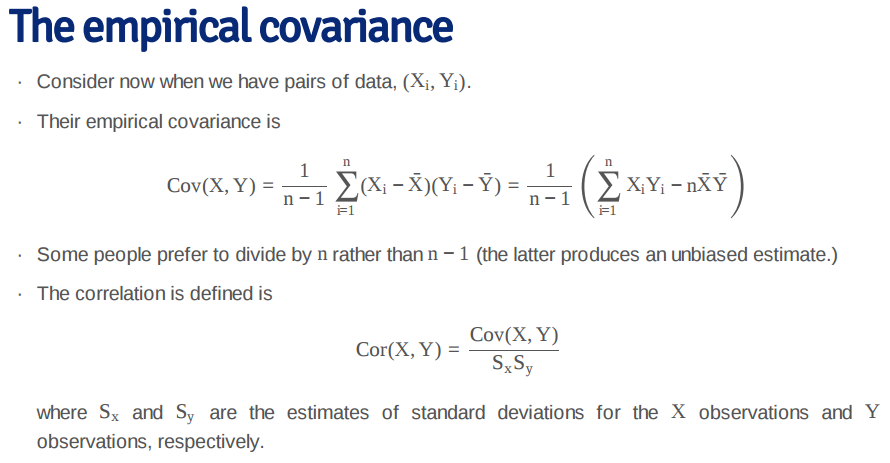
We can plot..  
**plot(galton$parent,galton$child,pch=19,col=”blue”)**

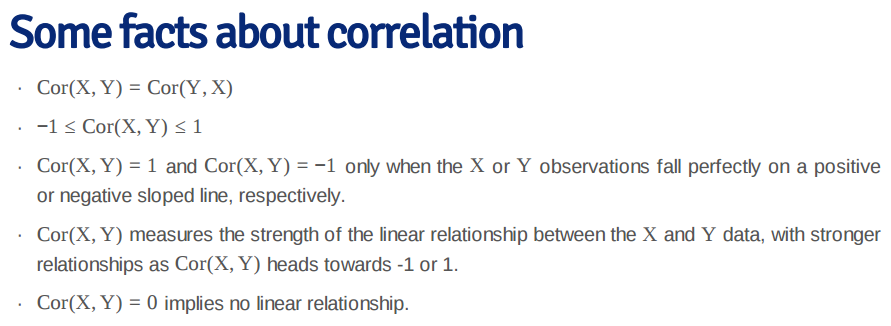
**Regression through the origin**  
-supposed that Xi are the parents’ heights  
-consider picking the slope beta that minimizes  
(Yi – XiB)^2  
-This is exactly using the origin as a pivot point picking the line that minimizes the sum of the squared vertical distances of the points of the line  
-use R studio’s manipulate function to experiment  
-subtract the means so that the original is the mean of the parent and children’s heights



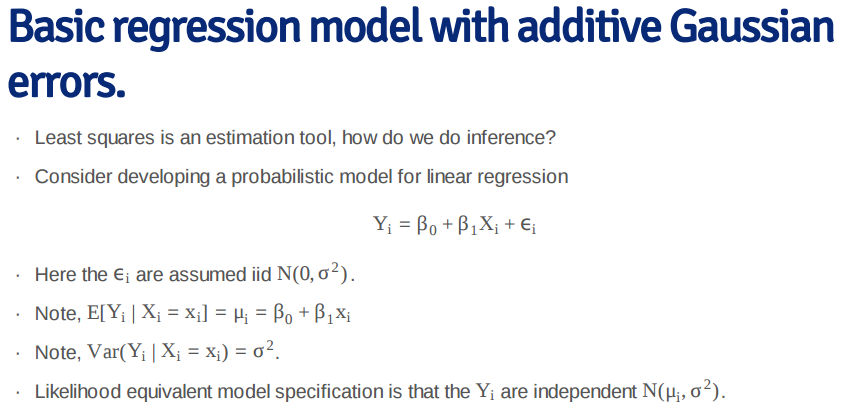
Visualizing the best fit line – size of points are frequencies at that X, Y combination

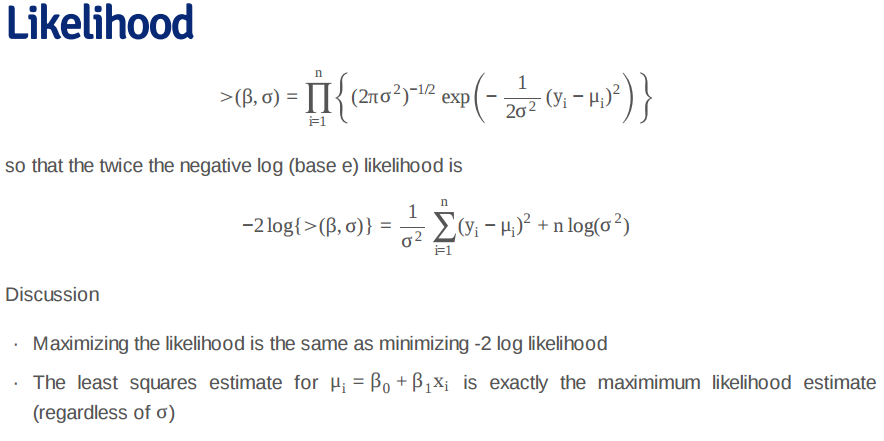
**Normalization**  
-the data defined by zi = (Xi – X) / s have empiricial mean zero and empirical stdev 1  
-the process of centering then scaling the data is called “normalizing” the data  
-normalized data are centered at 0 and have units equal to stdev of the original data  
-example, a value of 2 from normalized data means that data point was two stdev larger than the mean

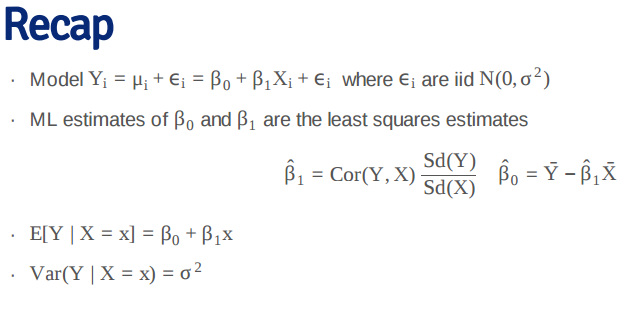


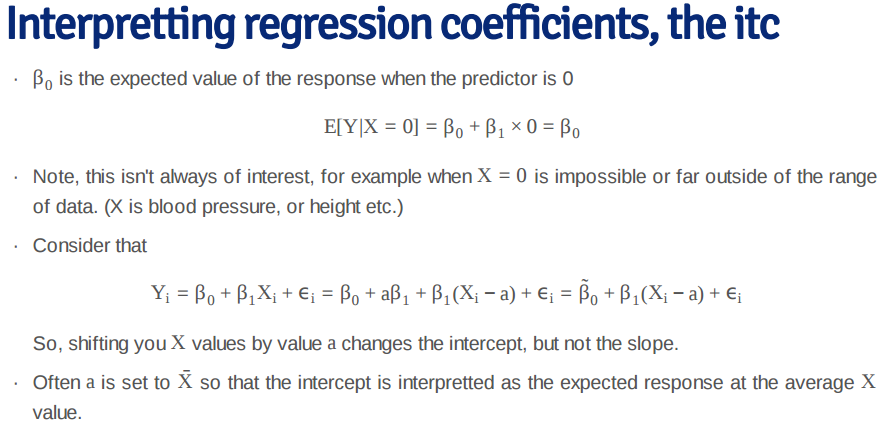
**Regression to the mean**-supposed that we normalize X (child’s height) and Y (parent’s height) so that they both have mean 0 and variance 1.  
-then recall our regression line passes through (0,0) (mean of X and Y)  
-if the slope of the regression line is Cor(Y, X), regardless of which variable is the outcome  
-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)

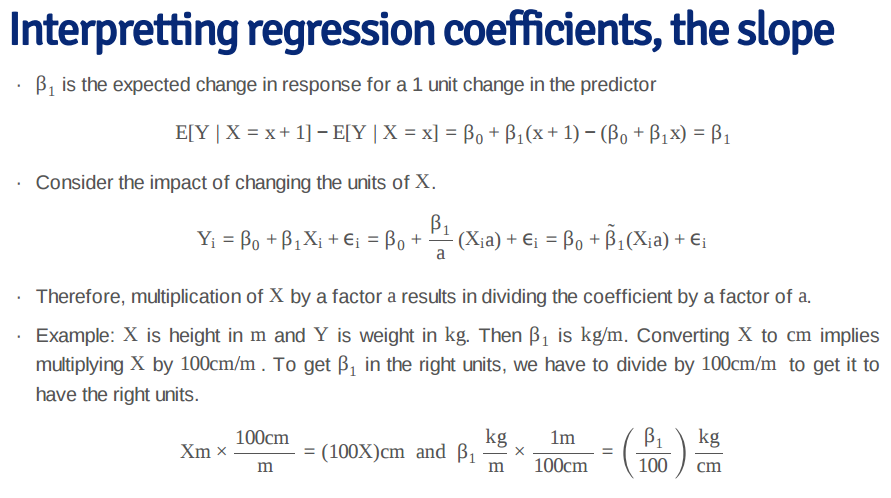
**Statistical Linear Regression Models**

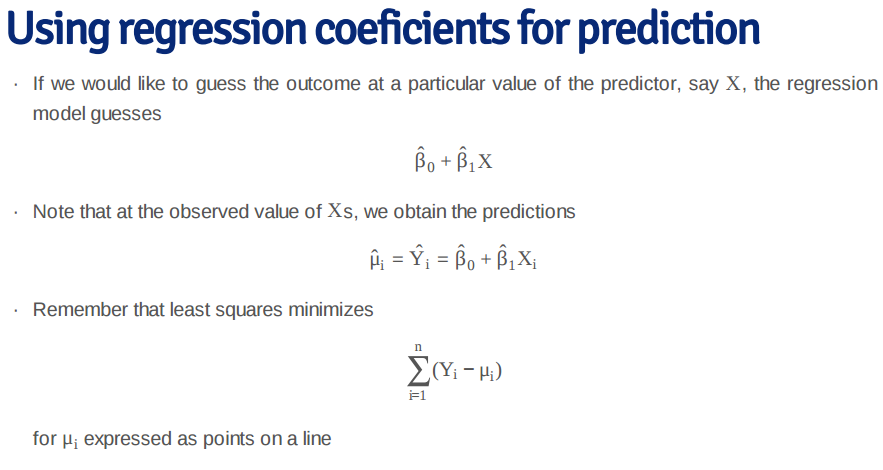








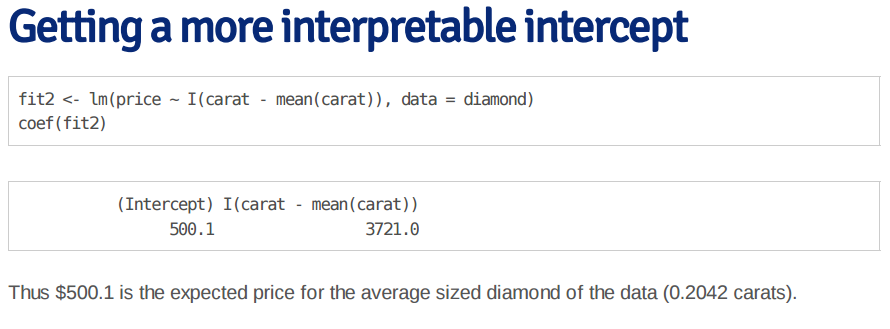




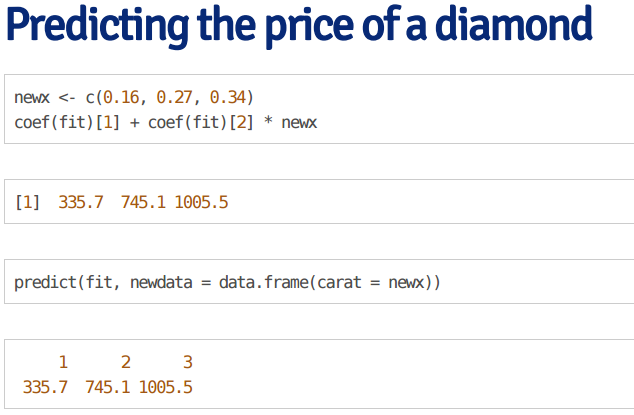
**Ex: Diamond data:**

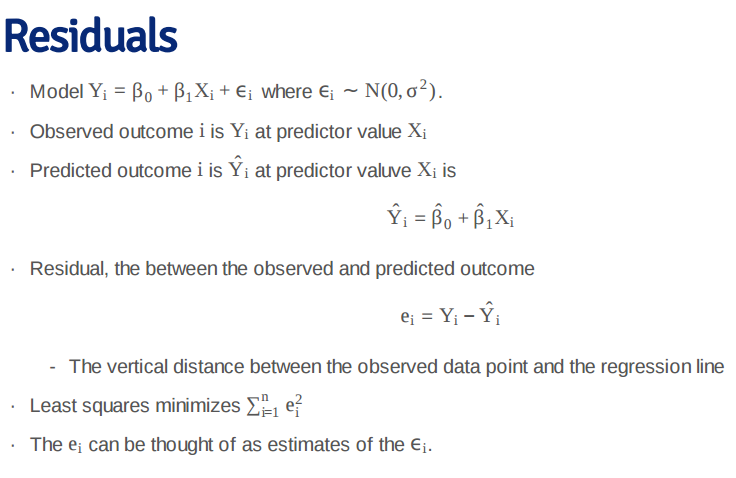
**Data(diamond)  
plot(diamond$carat, diamond$price,  
 xlab = “Mass (carats)”,  
 ylab = “Price (SIN $)”,  
 bg = “lightblue”,  
 col = “black”, cex = 1.1, pch = 21, frame = FALSE)  
 abline (lm(price ~ carat, data = diamond), lwd = 2)**

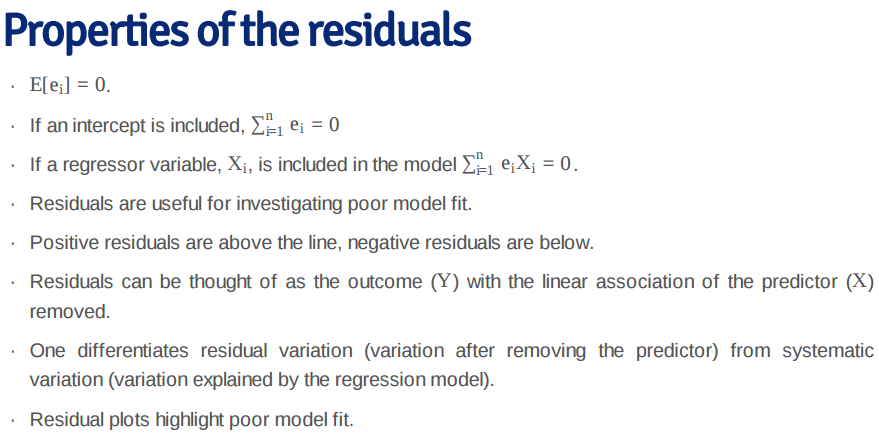


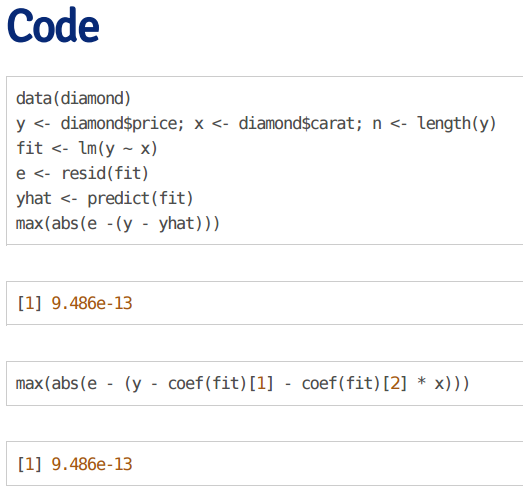


-a one carat increase in a diamond is pretty big, but we can also look at changes in 1/10th  
-if we just divide the coefficient by 10  
fit3 <- lm(price ~ I(carat \* 10), data = diamond)  
coef(fit3)









Non-linear data  
**x <- runif(100, -3, 3); y <- x + sin(x) + rnorm(100, sd = .2);  
plot(x, y); abline(lm(y ~ x))**

**Plot(x, resid(lm(y ~ x)));  
abline(h = 0)**

**Heteroskedasticity  
x <- runif(100, 0, 6); y <- x + rnorm(100, mean = 0, sd = 0.001 \* x);  
plot(x, y); abline(lm(y ~ x))**

Getting rid of blank space  
plot(x, resid(lm(y ~ x)));  
abline(h = 0)

