Residuals and residual variation

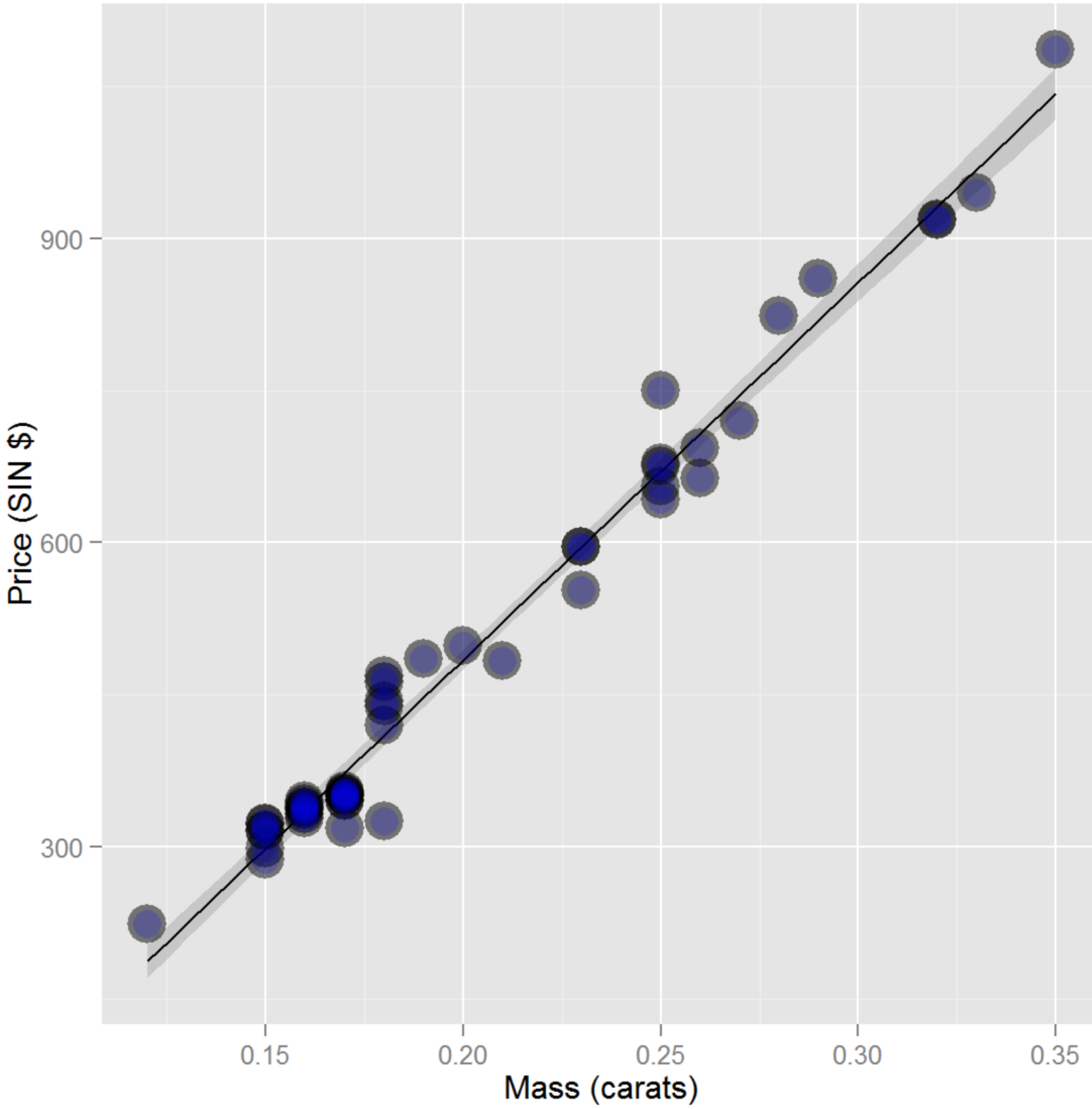
Brian Caffo, Jeff Leek and Roger Peng

Motivating example

`diamond` data set from `UsingR`

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)`

```
## Loading required package: MASS
## Loading required package: HistData
## Loading required package: Hmisc
## Loading required package: grid
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
## Loading required package: ggplot2
##
## Attaching package: 'Hmisc'
##
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
##
##
## Attaching package: 'UsingR'
##
## The following object is masked from 'package:ggplot2':
##
##     movies
##
## The following object is masked from 'package:survival':
##
##     cancer
```



Residuals

- Model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$.
- Observed outcome i is Y_i at predictor value X_i
- Predicted outcome i is $\hat Y_i$ at predictor valuve X_i is

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

- Residual, the between the observed and predicted outcome

$$e_i = Y_i - \hat Y_i$$

- The vertical distance between the observed data point and the regression line
- Least squares minimizes $\sum_{i=1}^n e_i^2$
- The e_i can be thought of as estimates of the ϵ_i.

Code

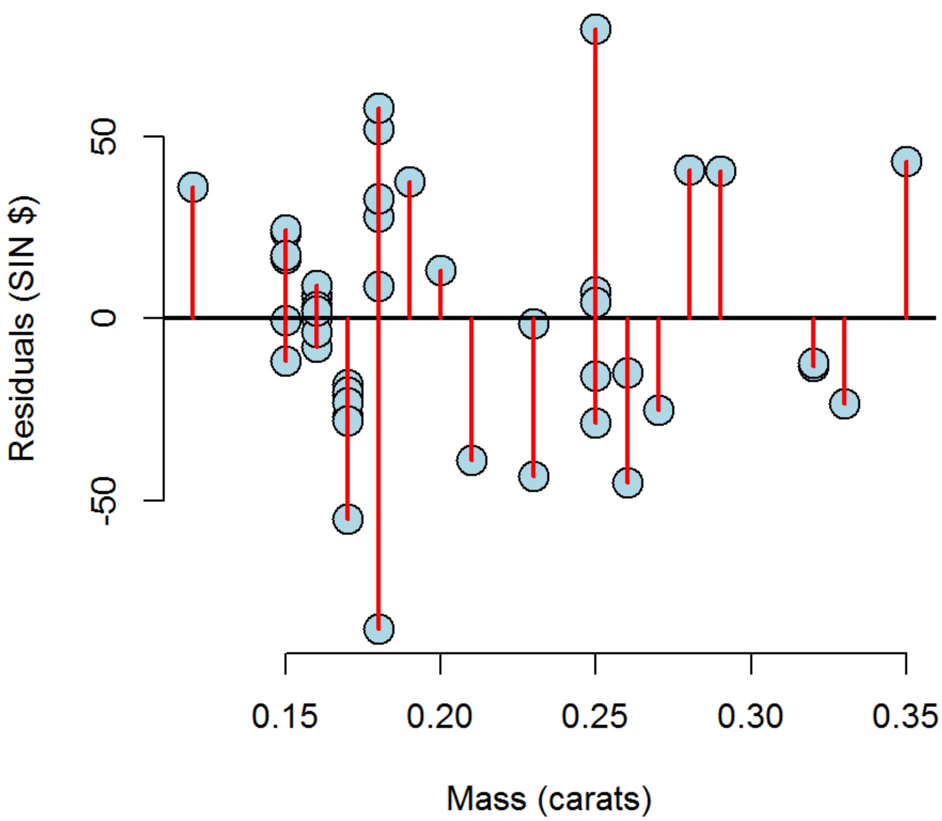
```
data(diamond)
y <- diamond$price; x <- diamond$carat; n <- length(y)
fit <- lm(y ~ x)
e <- resid(fit)
yhat <- predict(fit)
max(abs(e -(y - yhat)))
```

```
## [1] 9.485746e-13
```

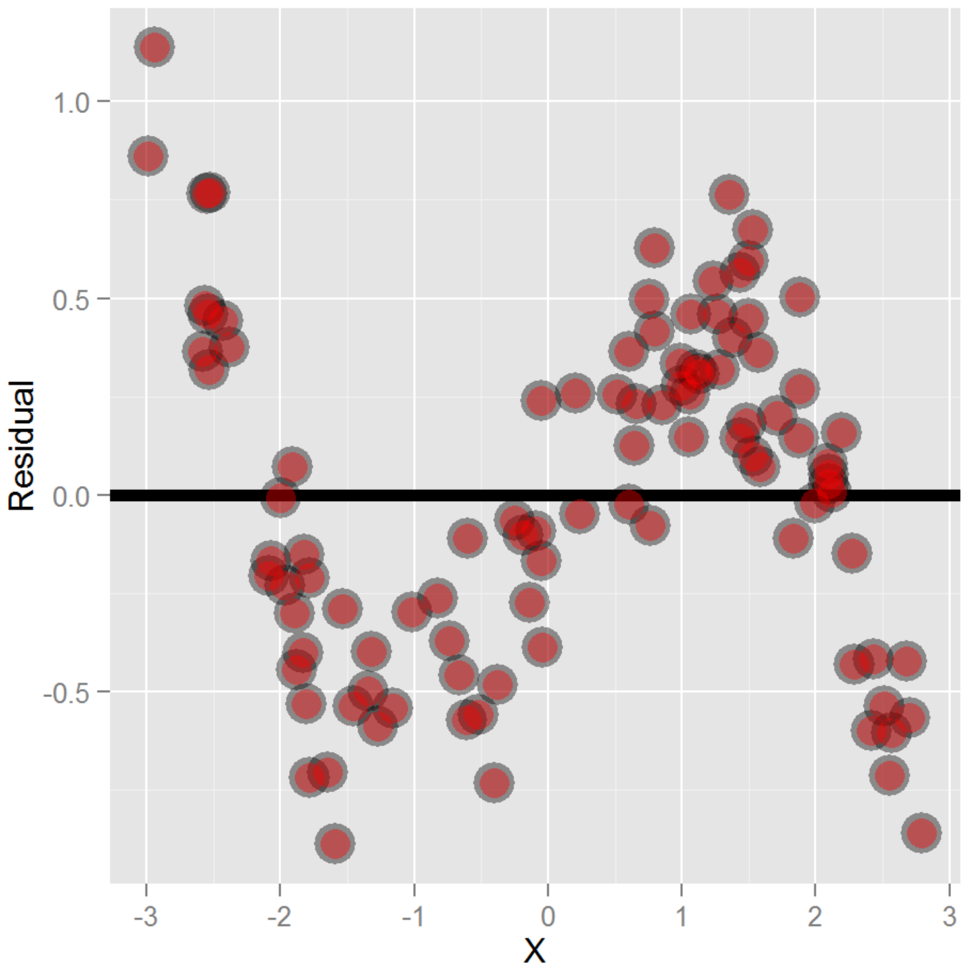
```
max(abs(e - (y - coef(fit)[1] - coef(fit)[2] * x)))
```

```
## [1] 9.485746e-13
```

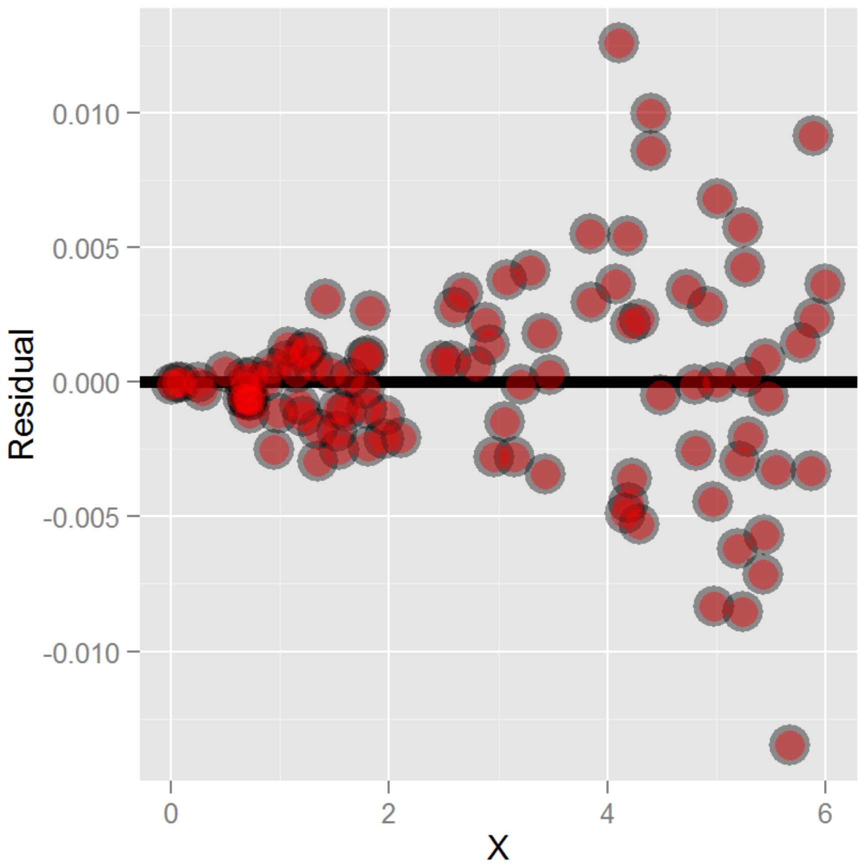
Residuals versus X



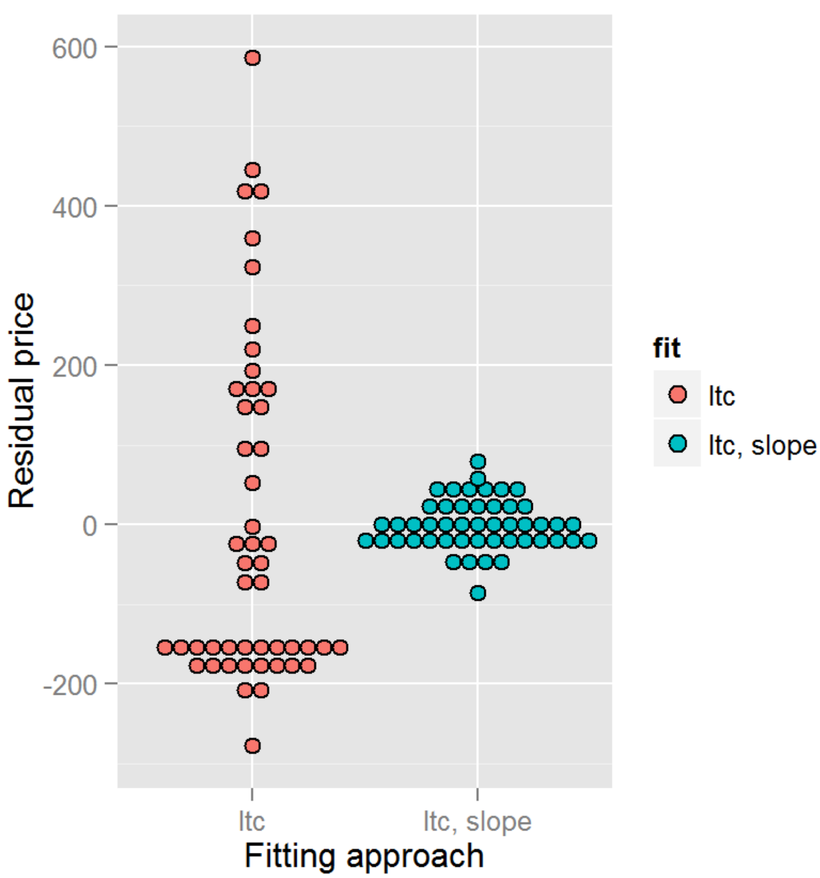
Residual plot



Getting rid of the blank space can be helpful



Diamond data residual plot



Diamond example

```
y <- diamond$price; x <- diamond$carat; n <- length(y)
fit <- lm(y ~ x)
summary(fit)$sigma
```

```
## [1] 31.84052
```

```
sqrt(sum(resid(fit)^2) / (n - 2))
```

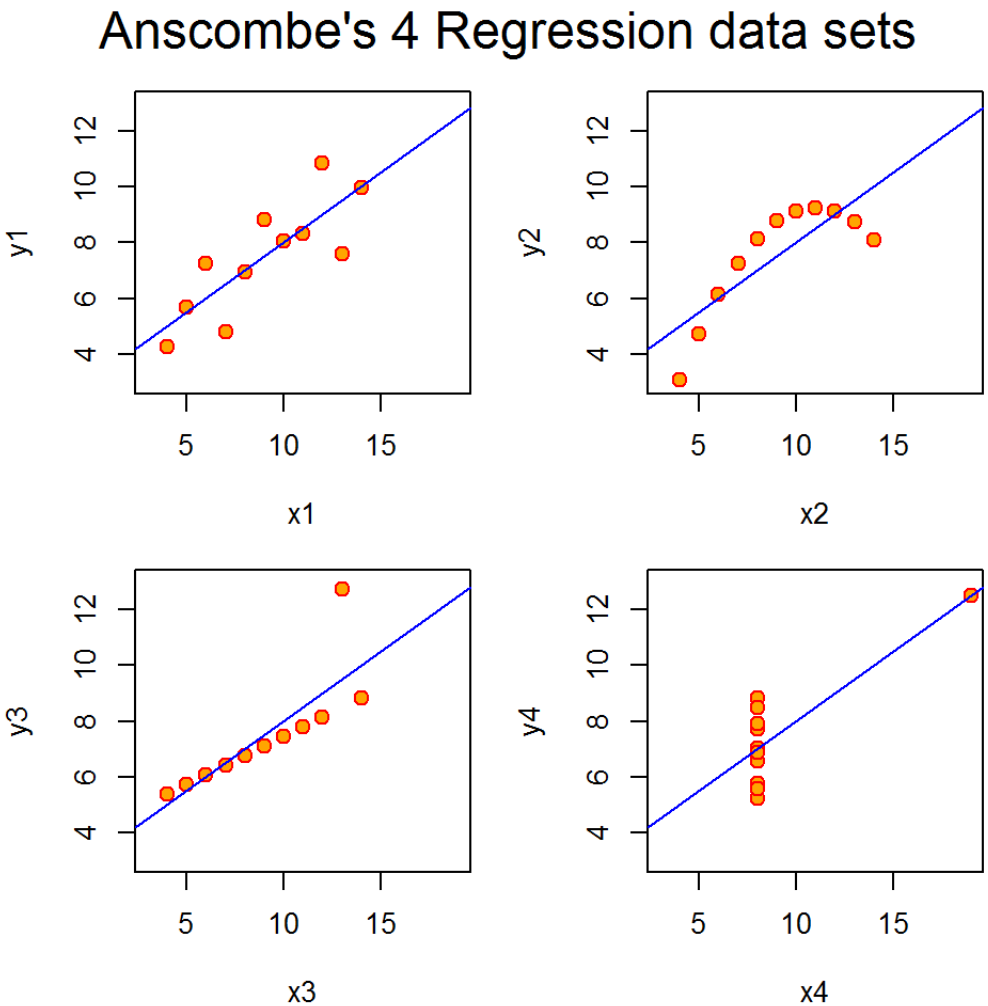
```
## [1] 31.84052
```

R squared

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

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

`data(anscombe);example(anscombe)`



The relation between R squared and r

(Again not required)

Recall that $(\hat{Y}_i - \bar{Y}) = \hat{\beta}_1 (X_i - \bar{X})$ so that

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \hat{\beta}_1^2 \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = Cor(Y, X)^2$$

Since, recall,

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

So, R^2 is literally r squared.