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.

Properties of the residuals

- $E[e_i] = 0$.
- If an intercept is included, $\sum_{i=1}^n e_i = 0$
- If a regressor variable, X_i, is included in the model $\sum_{i=1}^n e_i X_i = 0$.
- Residuals are useful for investigating poor model fit.
- Positive residuals are above the line, negative residuals are below.
- Residuals can be thought of as the outcome (Y) with the linear association of the predictor (X) removed.
- One differentiates residual variation (variation after removing the predictor) from systematic variation (variation explained by the regression model).
- Residual plots highlight poor model fit.

Code

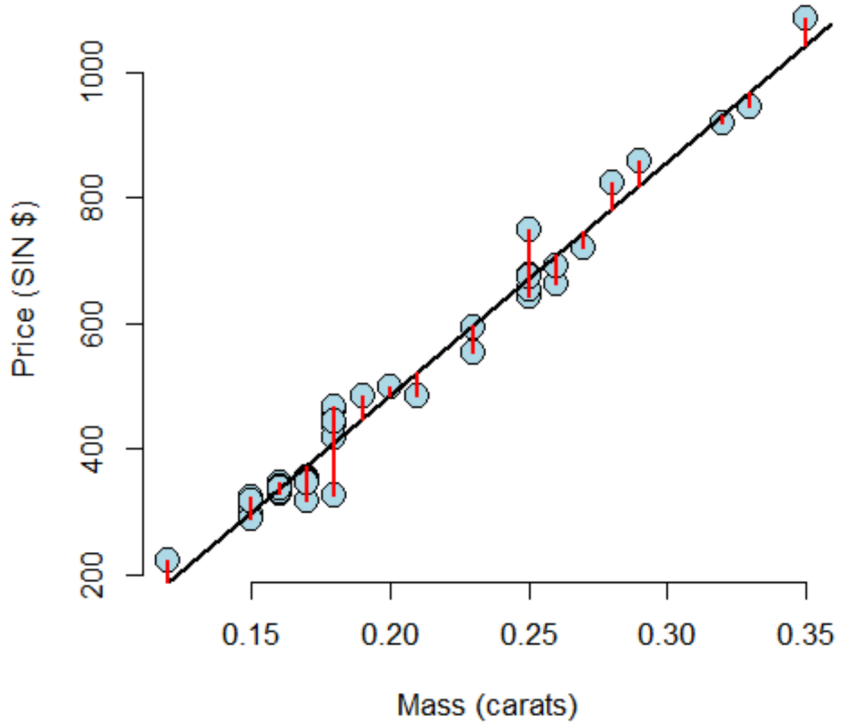
```
data(diamond)
y <- diamond$price; x <- diamond$carat; n <- length(y)
fit <- lm(y ~ x)
e <- resid(fit)
yhat <- predict(fit)
max(abs(e -(y - yhat)))
```

```
[1] 9.486e-13
```

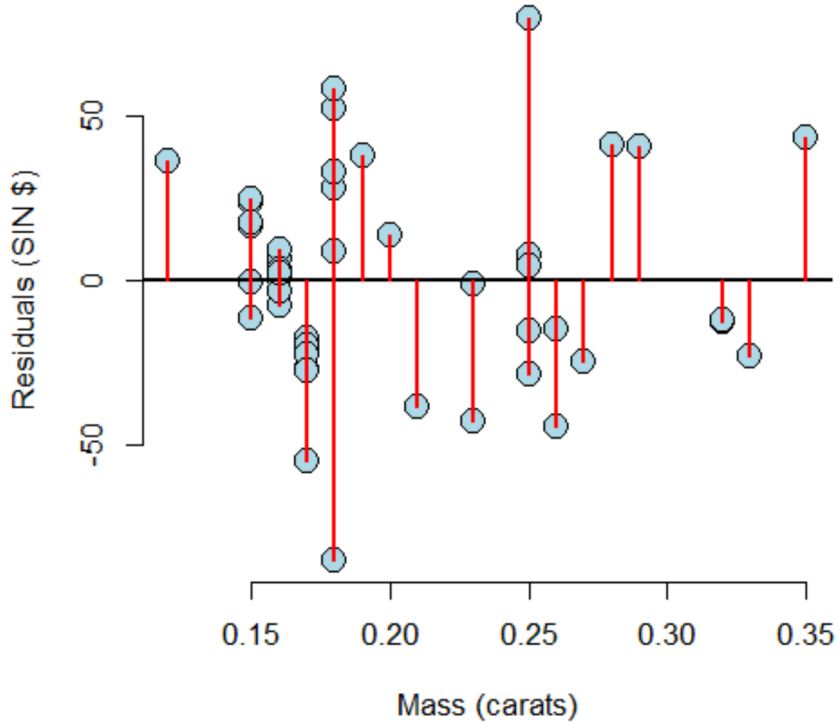
```
max(abs(e - (y - coef(fit)[1] - coef(fit)[2] * x)))
```

```
[1] 9.486e-13
```

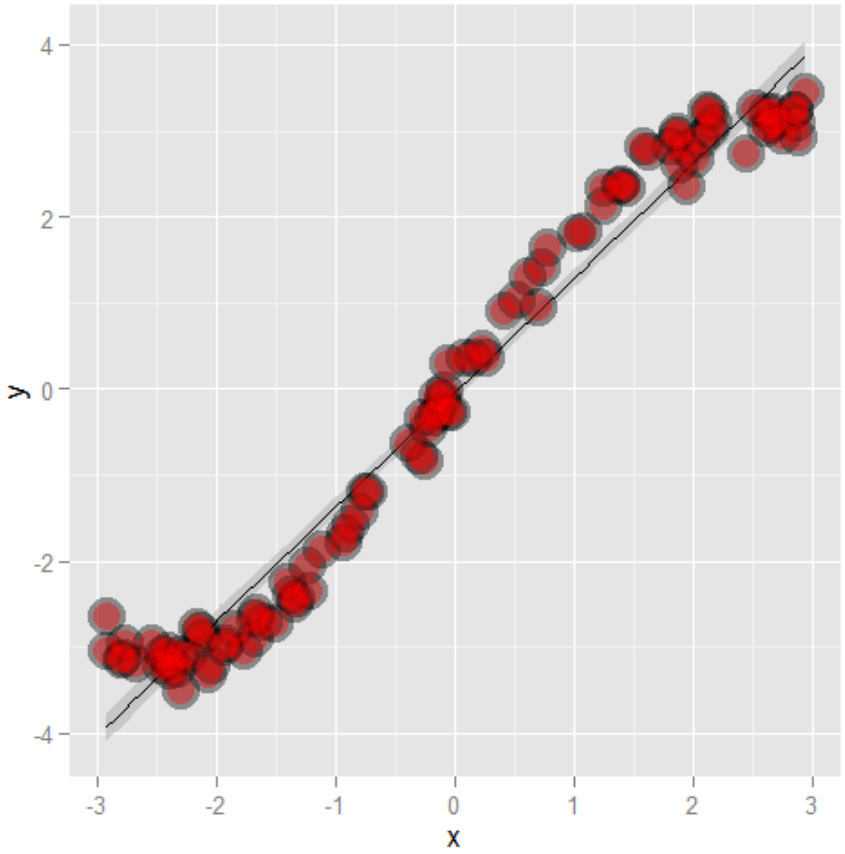
Residuals are the signed length of the red lines



Residuals versus X

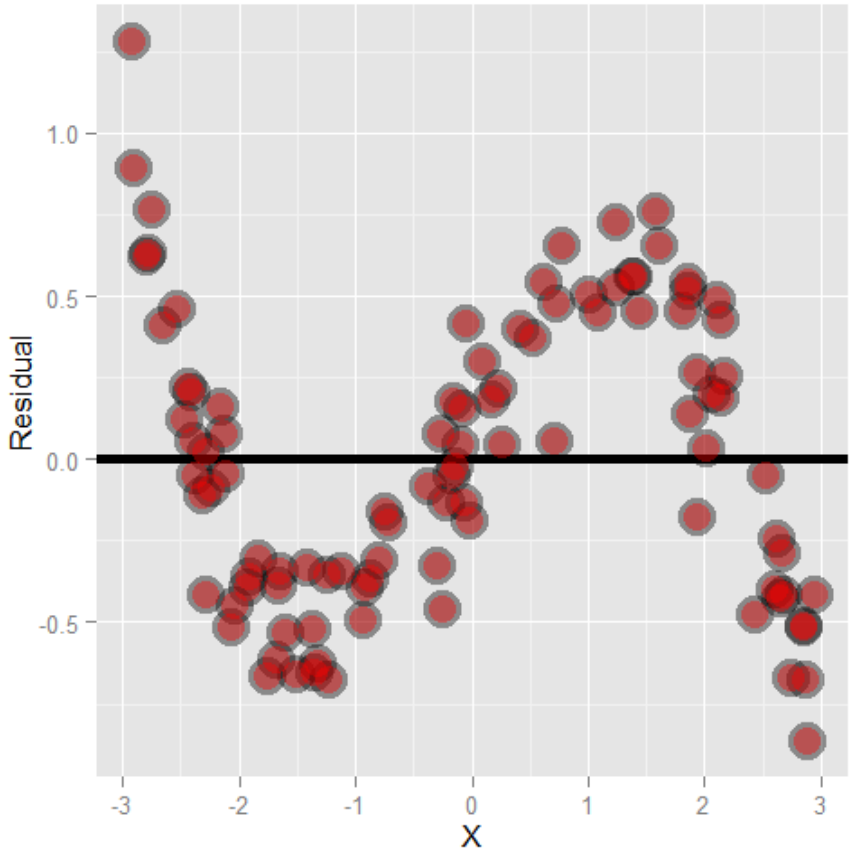


Non-linear data



9/22

Residual plot



10/22

Heteroskedasticity