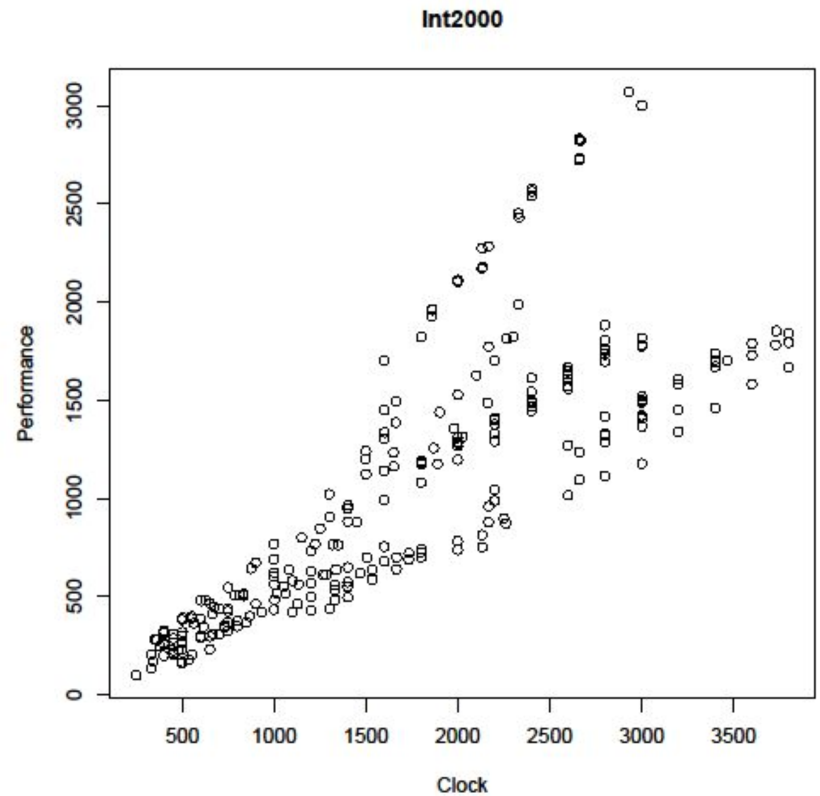
Lab 3: Simple Linear Regression (SLR)

- $\hat{y} = a_0 + a_1 x_1$
- $\hat{y}$ = output
- The ^ means predicted or estimated value, not an actual observed value.
- a_0 and a_1 = regression coefficients
- x_1 = input data
- Given many x_1 and y pairs, our goal is to compute a_0 and a_1

Does it look linear?

- plot() function

plot(int00.dat[,"clock"],
int00.dat[,"perf"],
main="Int2000",
xlab="Clock",
ylab="Performance")



Linear model function

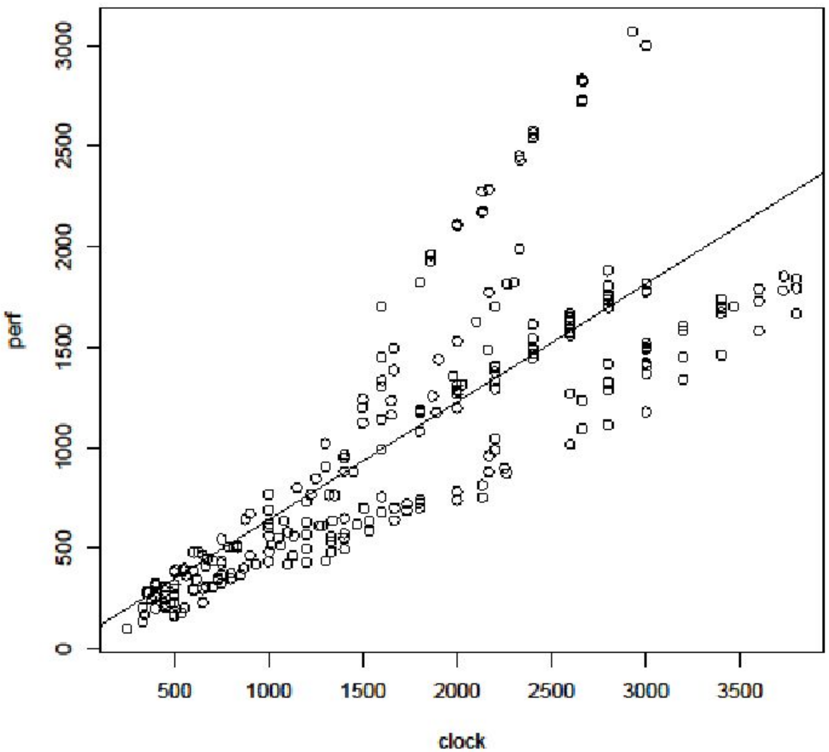
- lm()
- int00.lm <- lm (perf ~ clock)
- $\hat{y} = a_0 + a_1 x_1$
- $\hat{y}$ = perf
- x_1 = clock
- *See example in Section 3.2*
- The ~ indicates a relationship between the two variables – read as "by"
  - "Model a linear relationship for perf by clock"

Linear model function

- lm()
- int00.lm <- lm (perf ~ clock)
- $\hat{y} = a_0 + a_1 x_1$
- $\hat{y}$ = perf
- x_1 = clock
- *See example in Section 3.2*
- $a_0 = 51.8$
- $a_1 = 0.586$
- *Final model:* perf = 51.8 + 0.586 * clock

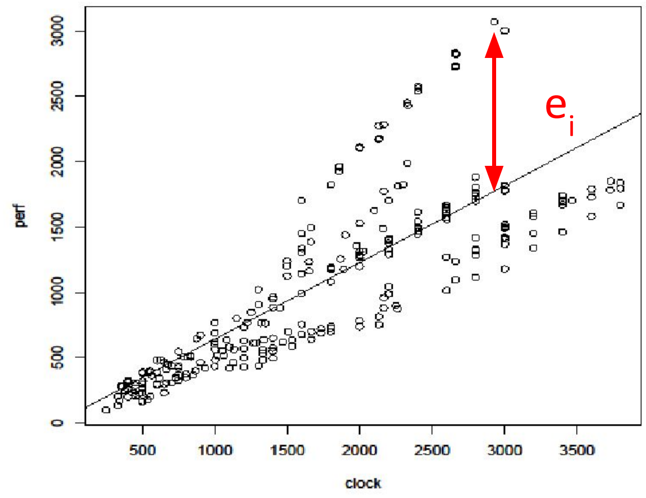
Plotting the regression line

- plot(clock,perf)
- abline(int00.lm)



Behind the Code

- For i=1 to n (x_i, y_i) pairs
- $y_i = a_0 + a_1 x_1 + e_i$
- e_i = *residual*
  = (predicted by line) - (actual)



- Minimize sum-of-squares of residuals

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

Behind the Code

- Set partial derivatives to zero to find minima

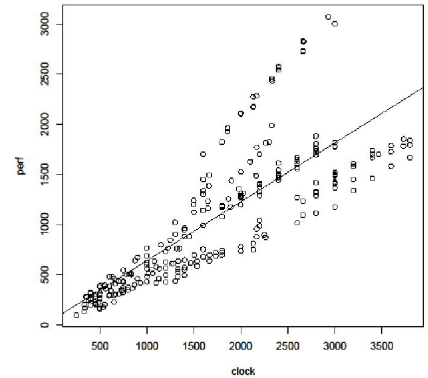
$$\frac{\partial SSE}{\partial a_0} = 0, \frac{\partial SSE}{\partial a_1} = 0$$

$$na_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

- Two equations, two unknowns
  – Solve for a_0, a_1

Model quality



- Section 3.3
- summary(int00.lm)
- Residuals = actual value – fitted value
  - above/below line = positive/negative residual
  - Expect median of residuals ≈ 0
  - Expect residuals normally distributed around 0
  - Expect min, max approx same distance from 0

```
Residuals:
Min        1Q     Median   3Q    Max
-634.61 -276.17 -30.83 75.38 1299.52
```

Model quality – coefficients

- Standard error = measure of total variation in residuals
  - If normally distributed expect std err = 5-10x smaller than coefficient value
  - Small value small variability
- Pr(>|t|) = p-value = Pr(you obtain a t-value this large or larger if null hypothesis is true)
  - Tests *null hypothesis* that this variable has no correlation with dependent variable, i.e., this variable has zero effect on output
  - If p-value < significance level (α), sample data is sufficient to reject the null hypothesis for population
  - If (p-value < α) coefficient has higher probability of non-zero effect *strong evidence* it is probably useful in the model
  - Typically choose α = 0.05 or 0.1

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 51.78709 53.31513    0.971    0.332
clock        0.58635  0.02697   21.741   <2e-16 ***
```

Model quality – visual indicators

- ***
  - $0 < p \leq 0.001$
- **
  - $0.001 < p \leq 0.01$
- *
  - $0.01 < p \leq 0.05$
- .
  - $0.05 < p \leq 0.1$
- [blank]
  - $0.1 < p \leq 1$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)51.78709 53.31513 0.971 0.332
clock       0.58635  0.02697 21.741 <2e-16 ***
```

- Quick visual indicator of the significance of that coefficient.

Model quality – residuals

- Residual standard error
  - Measure of total variation in residual values
  - If residuals are normally distributed
    - Expect 1st and 3rd quantiles ≈ 1.5 * standard error

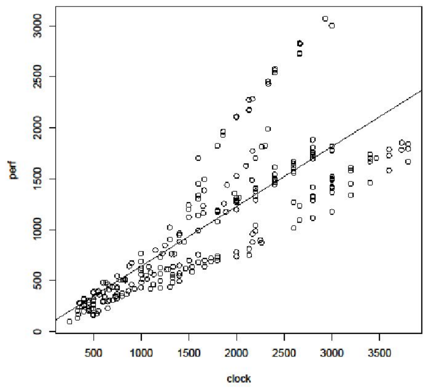
```
Residuals:
Min        1Q    Median   3Q   Max
-634.61 -276.17 -30.83 75.38 1299.52


Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept)51.78709 53.31513 0.971 0.332
clock      0.58635  0.02697 21.741 <2e-16 ***
```

Model quality – R^2 and F-statistic

- Degrees of freedom = (number of observations used to generate model) – (number of coefficients)
  - E.g. 256 observations used to compute 2 coefficients
    - 254 degrees of freedom
- R^2 = percentage of total variation explained by the model

  $0 < R^2 < 1$
- Adjusted R^2 – discuss later for multi-factor models
- F-statistic
  - Compares current model to a model that has only the intercept parameter
  - Discussed further in Chapter 4

R^2 – A Little Deeper



$$SST = \sum (y_i - \bar{y})^2$$

$$SSE = \sum (y_i - a_0 - a_1 x_i)^2$$

$$SSR = SST - SSE = a_1 \sum (x_i - \bar{x})^2 (y_i - \bar{y})^2$$

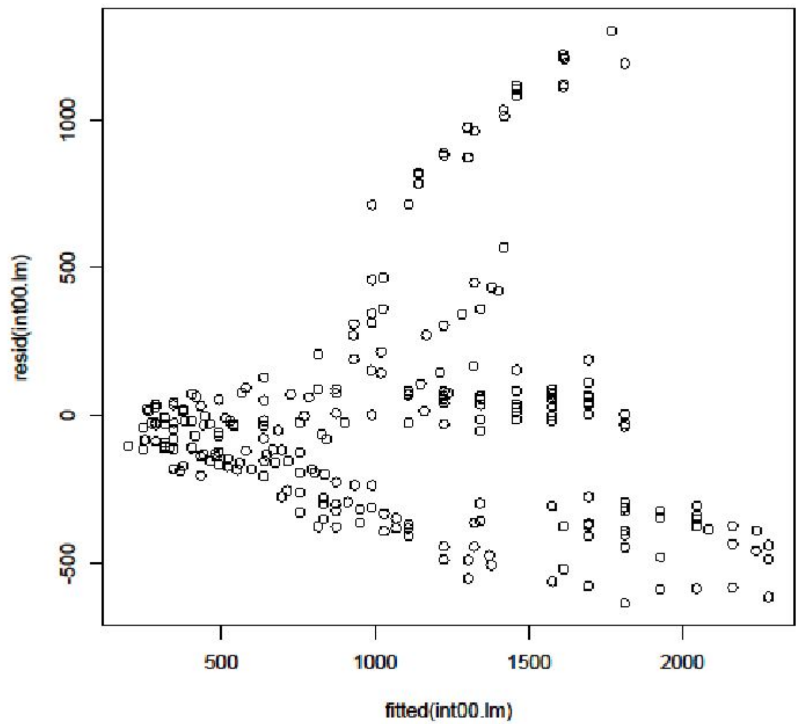
$$SST = SSR + SSE$$

$$R^2 = SSR / SST$$

- R^2 = coefficient of determination
  - = Fraction of total variation explained by model

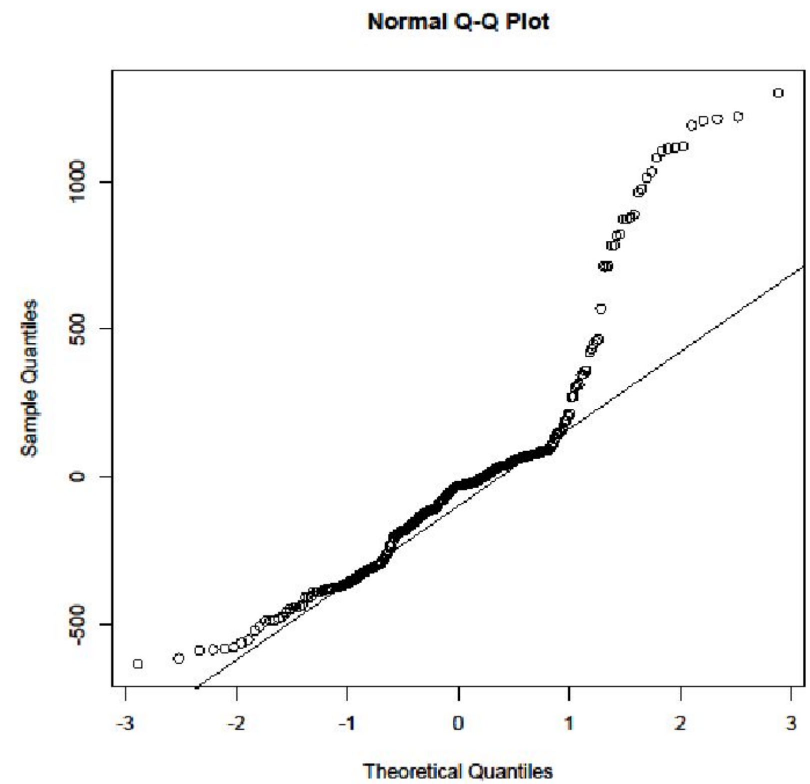
Residual analysis

- Section 3.4
- $y = a_0 + a_1 x_1 = f(x_i)$
- fitted(int00.lm) computes
  $y_f = f(x_i)$ for all x_i
- resid(int00.lm) computes
  $y_i - y_f$ for all outputs y_i
  Residuals = (actual value) – (fitted value)
- plot(fitted(int00.lm), resid(int00.lm))
  - Expect randomly distributed around 0
- Note vertical lines
  - Different perf (y_i) with same clock (x_i)



Residual analysis – QQ plot

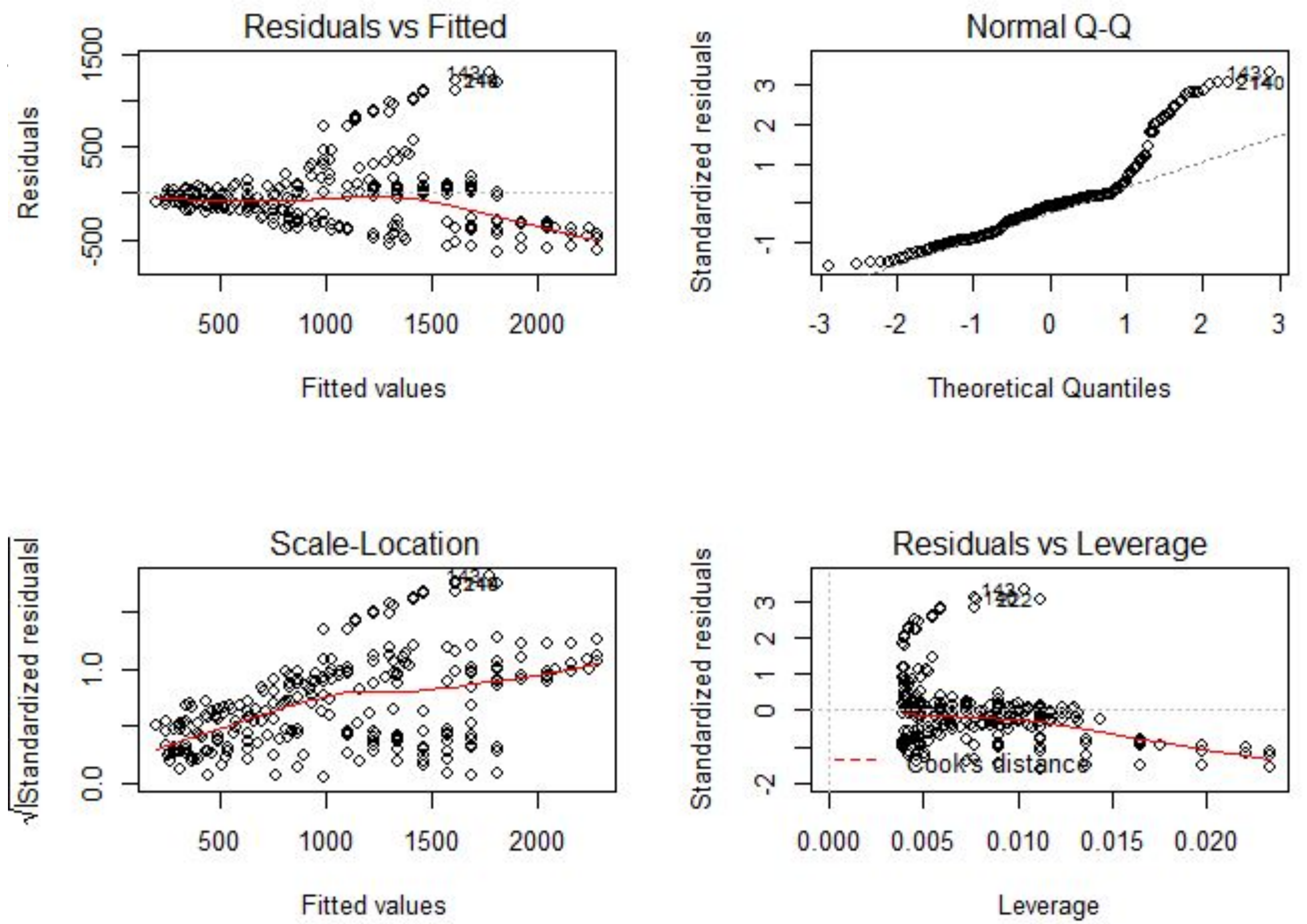
- Quantile vs quantile (QQ) plot
  - qqnorm(resid(int00.lm))
  - qqline(resid(int00.lm))
  - Expect the points to follow a straight line if error is normally distributed



Both the residual plot and QQ plot suggest that a one-factor model is insufficient for this data, but it still may be useful.

Simpler Generation of Diagnostic Plots

```
par(mfrow=c(2,2))
plot(int00.lm)
```



To do

- Read Chapter 3
- Download and complete Lab 3