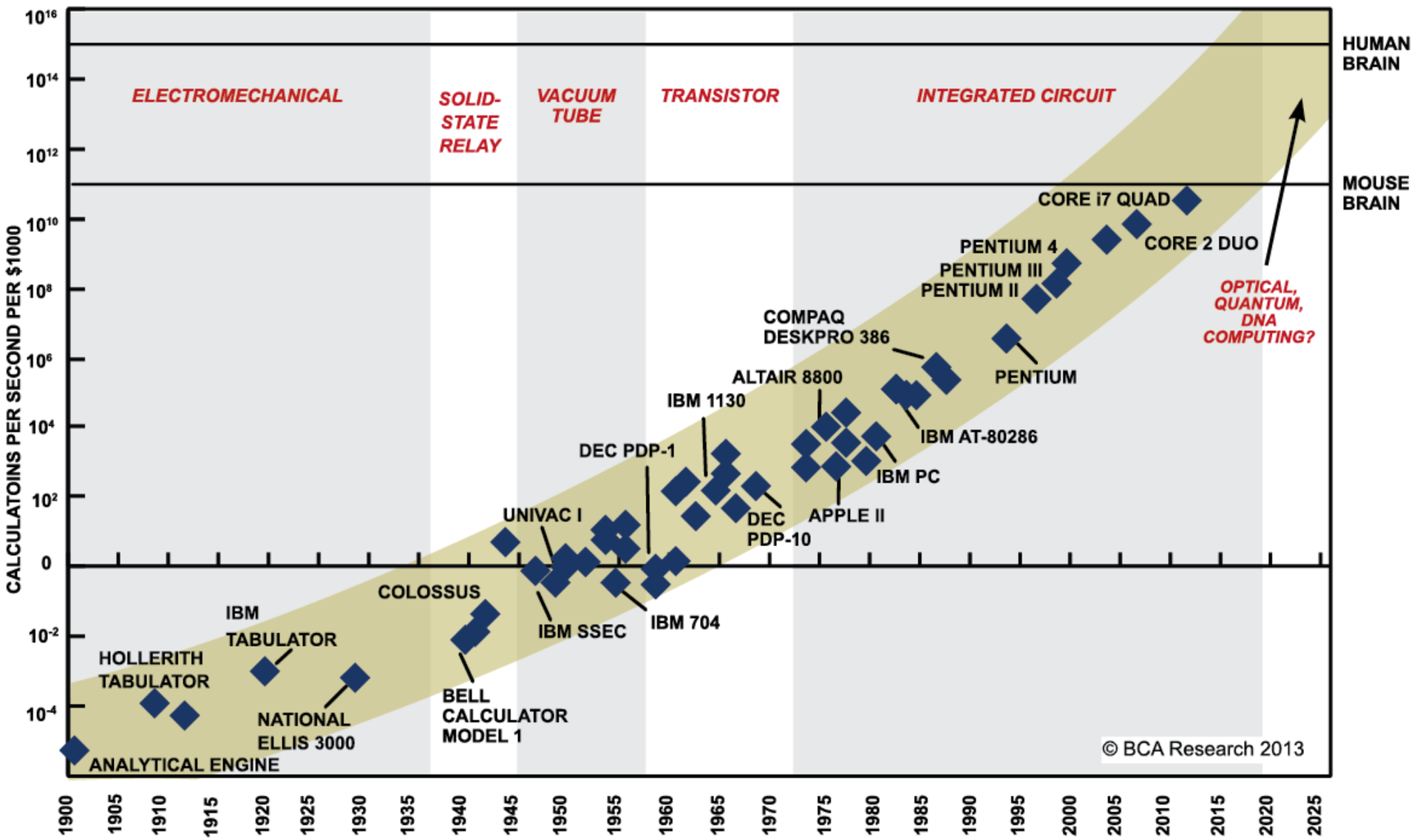
Multiple Regression

- Example, Moore's Law: In 2014 Google's CEO Eric Schmidt discussed the future of the internet and said that according to Moore's Law, in 10 years every computer device you use will be 100 times cheaper or 100 times faster"



SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

Multiple Regression

- In 1965 Moore predicted that the number of transistors would double every year and in 1975 he modified that doubling time to every two years. If the rate of computing power is assumed constant, Moore's Law looks like:

inverse of the time to double

$$y = b_0 2^{b_1 t}, \quad t \geq 0,$$

measurement at time t

initial measurement

Taking log base 2 we have:

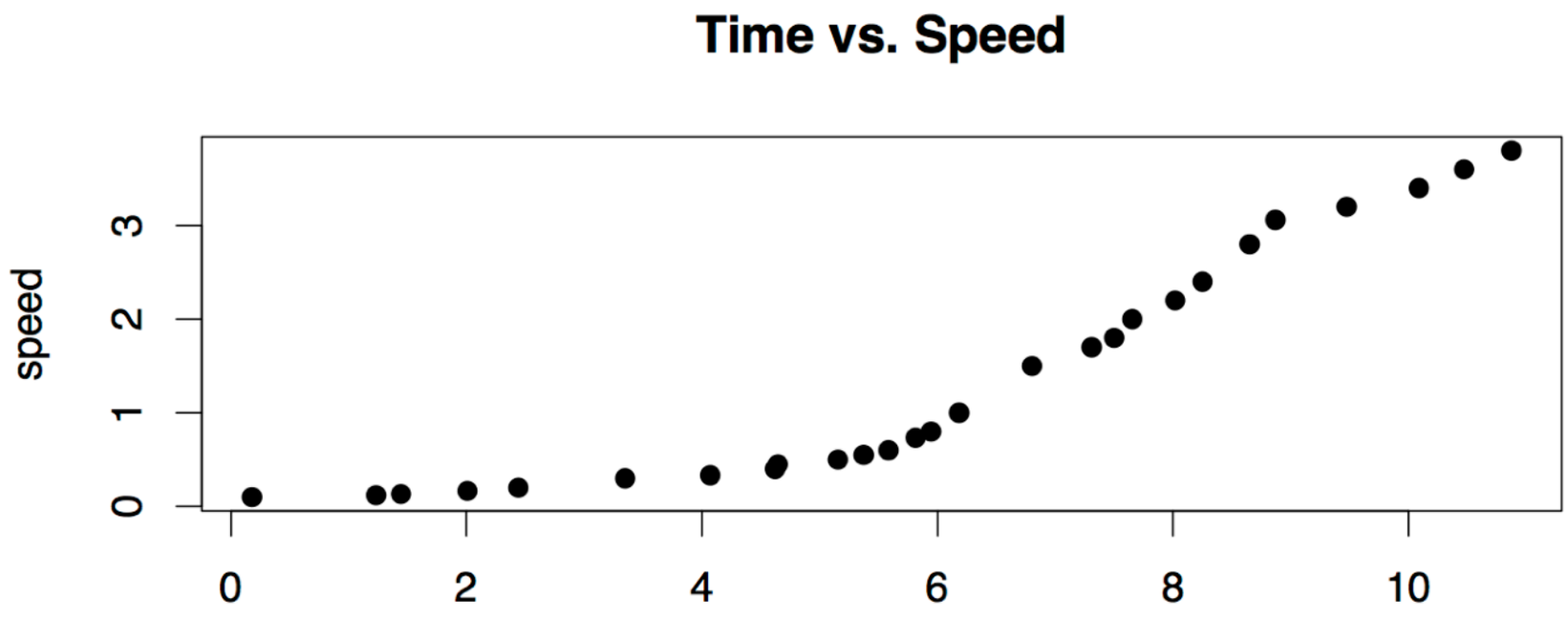
$$\log_2(y) = \log_2(b_0) + b_1 t, \quad t \geq 0$$

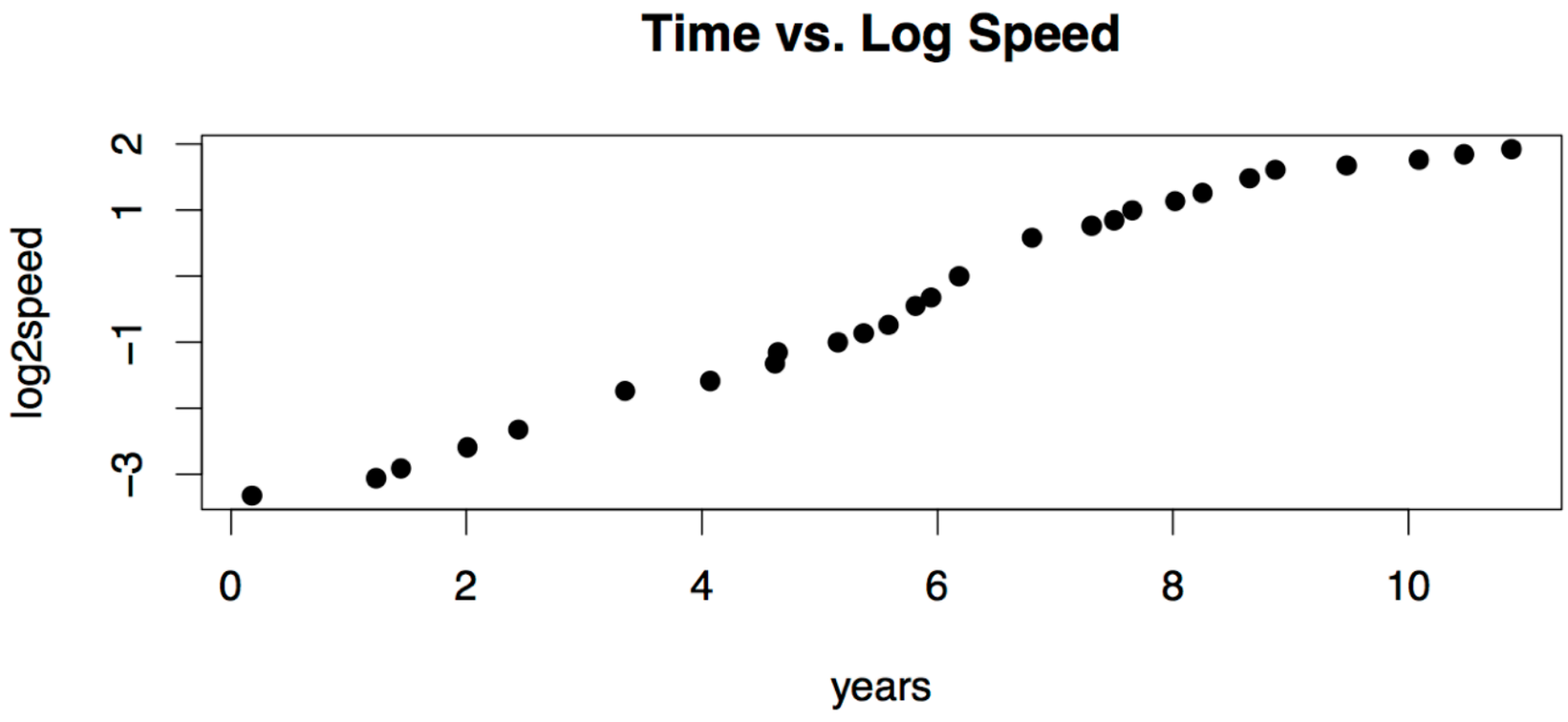
Multiple Regression

- Data:

```
> CPUspeed = read.table("CPUspeed.txt", header=TRUE)
> head(CPUspeed)
  year month day      time speed log10speed
1 1994     3   7 1994.179 0.100 -1.0000000
2 1995     3  27 1995.233 0.120 -0.9208188
3 1995     6  12 1995.444 0.133 -0.8761484
4 1996     1   4 1996.008 0.166 -0.7798919
5 1996     6  10 1996.441 0.200 -0.6989700
6 1997     5   7 1997.347 0.300 -0.5228787


> years = CPUspeed$time - 1994
> speed = CPUspeed$speed
> log2speed = CPUspeed$log10speed / log10(2)
```

Multiple Regression



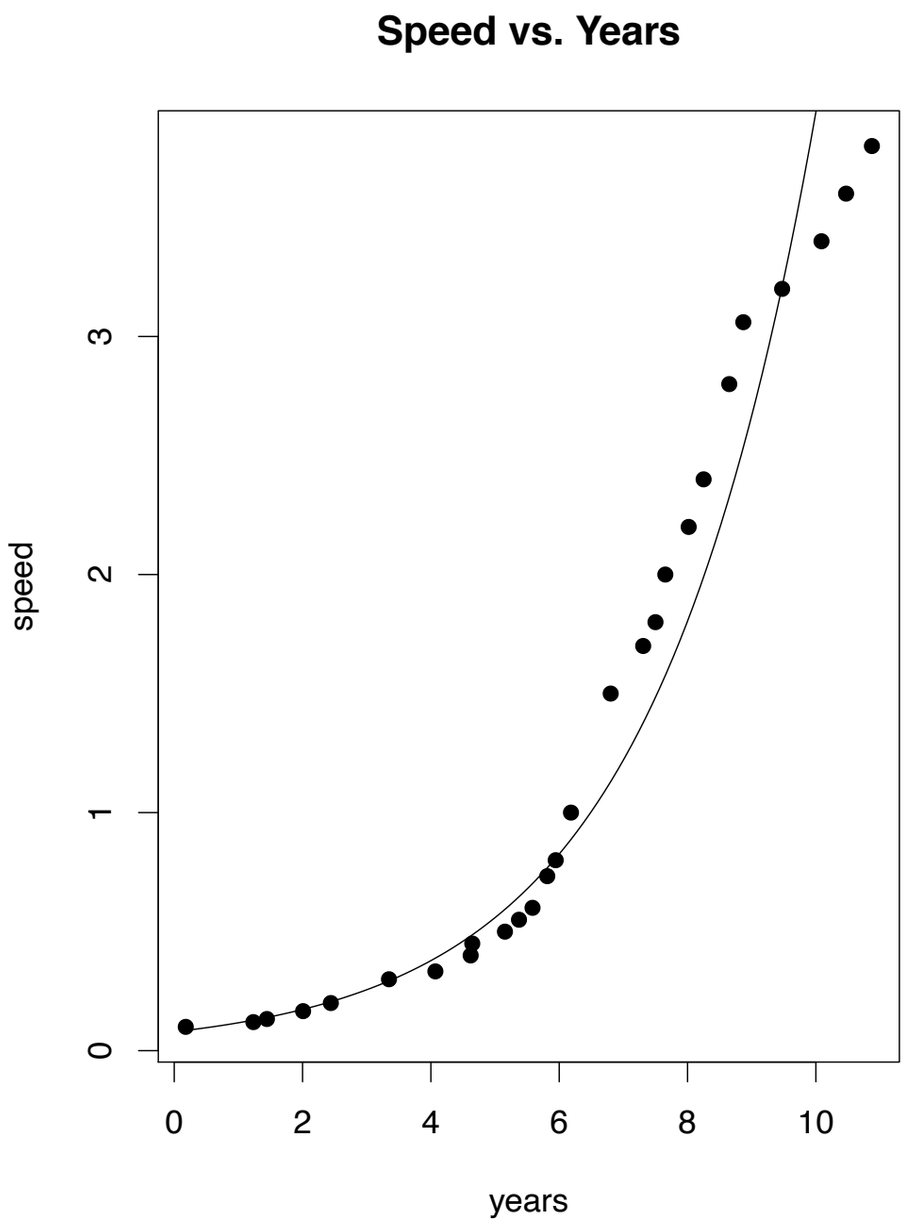


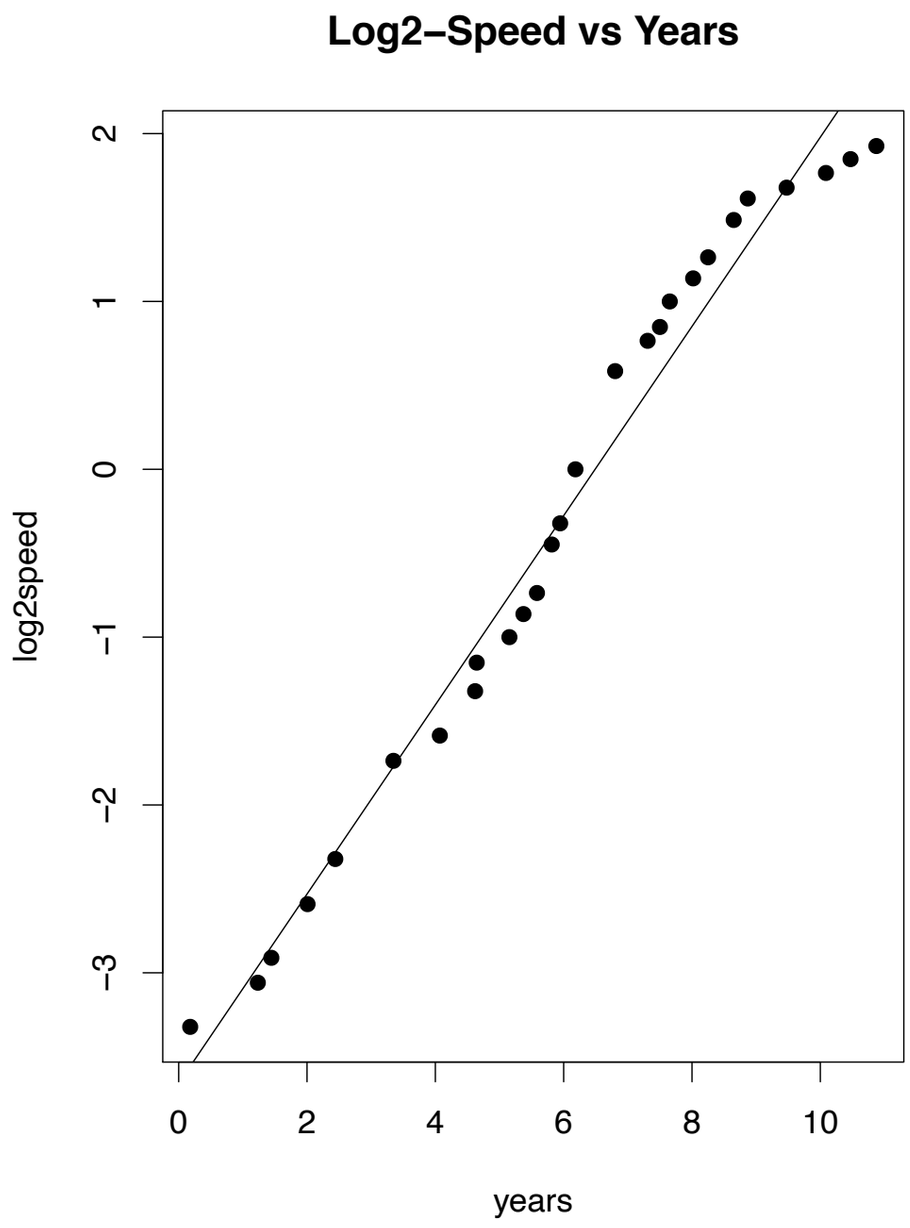
Multiple Regression

```
> L = lm(log2speed ~ years)
> summary(L)

Call:
lm(formula = log2speed ~ years)

Residuals:
     Min       1Q   Median       3Q      Max
-0.54486 -0.22429 -0.03959  0.26914  0.40891

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.65814    0.11499  -31.81   <2e-16 ***
years        0.56367    0.01726   32.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2595 on 25 degrees of freedom
Multiple R-squared:  0.9771, Adjusted R-squared:  0.9762
F-statistic:  1066 on 1 and 25 DF,  p-value: < 2.2e-16
```

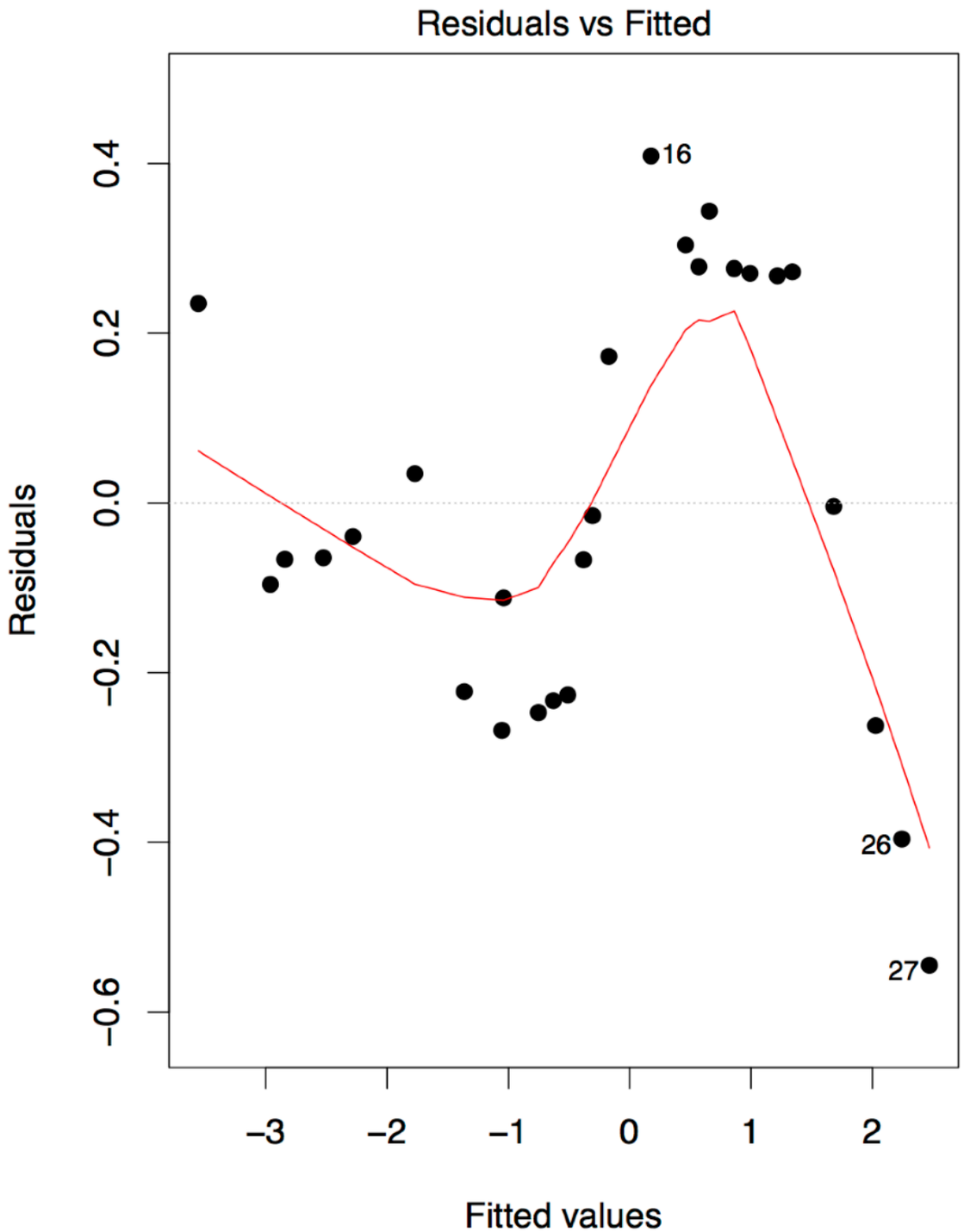
Multiple Regression

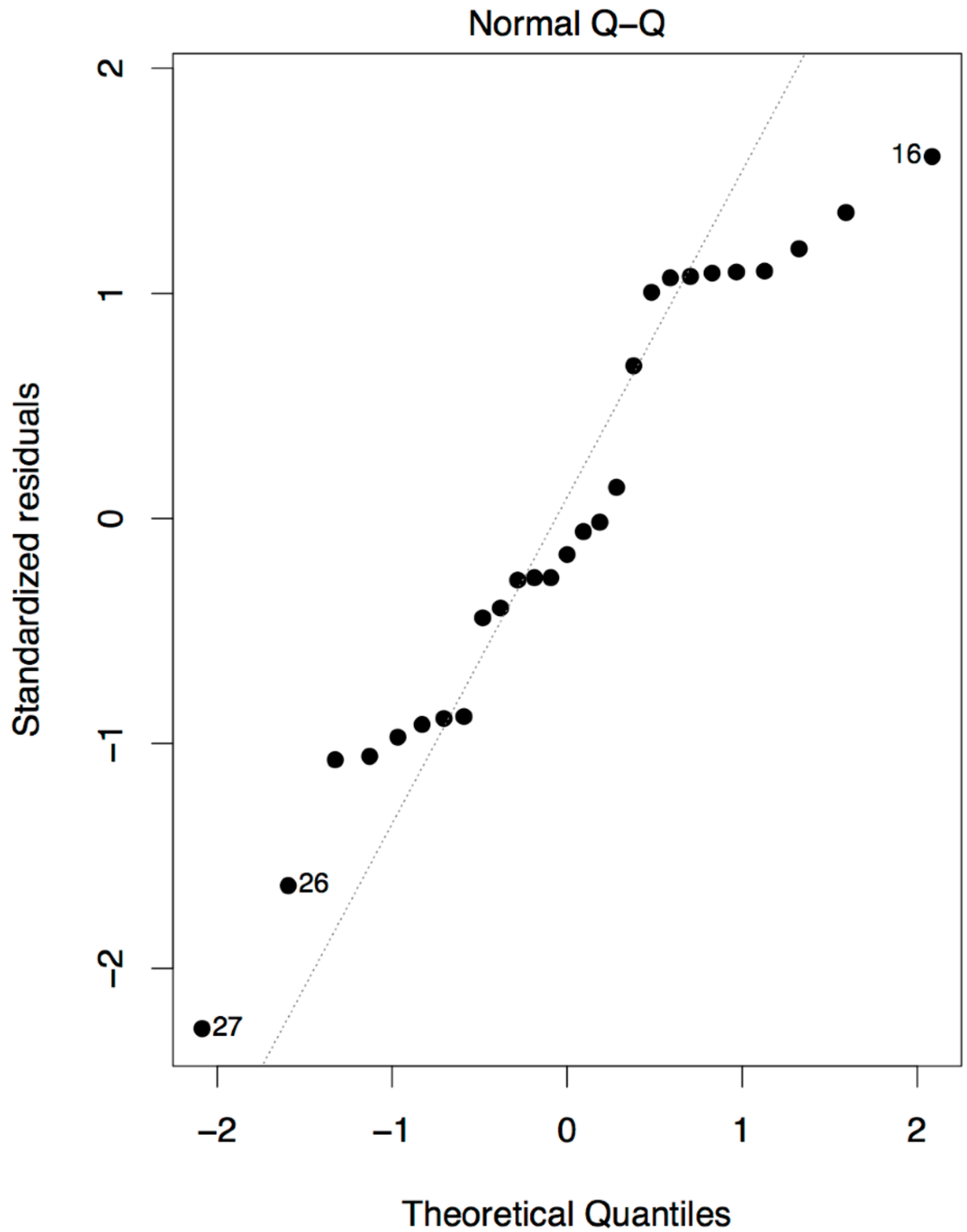
```
> plot(years, speed,pch=19,main="Speed vs. Years")
> curve(2^(-3.6581 + 0.5637 * x), add=TRUE)
> plot(years, log2speed,pch=19,main="Log2-Speed vs Years")
> abline(L)
```





Multiple Regression





Multiple Regression

- Comparing non-nested models: AIC, BIC, coefficient of determination (adjusted). Key functions in **R:** `extractAIC, AIC, BIC, logLik`

- **Variable selection:** Choose "the best" subset of predictors. Why?
  - Explain the data in the simplest way, remove redundant predictors. Occam's Razor: among several explanations for a given phenomenon, the simplest is the best
  - Unnecessary predictors will add noise in estimation
  - Collinearity
  - Cost: Avoid measuring unnecessary predictors

Before variable selection: Consider transformations, identify outliers and influential points

Multiple Regression

Stepwise procedures:

- Backward elimination: begin with all predictors, remove the predictor with the highest p-value greater than a threshold, refit model and do the same until all p-values are less than threshold. We could also drop predictors based on AIC: `drop1`

- Forward selection: begin with no predictors, add predictors based on p-values (begin with one with the lowest p-value smaller than a threshold and continue). We could also add predictors based on AIC: `add1`

- Stepwise regression: Combination of backward elimination and forward selection, e.g., function `step`

Multiple Regression

Example: The data frame `swiss` contains standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888. The variables are Fertility, Education, Agriculture, Examination, Education, Catholic, and Infant.Mortality

```
> head(swiss)
             Fertility Agriculture Examination Education Catholic Infant.Mortality
Courtelary        80.2        17.0          15        12     9.96             22.2
Delemont          83.1        45.1           6         9    84.84             22.2
Franches-Mnt      92.5        39.7           5         5    93.40             20.2
Moutier           85.8        36.5          12         7    33.77             20.3
Neuveville        76.9        43.5          17        15     5.16             20.6
Porrentruy        76.1        35.3           9         7    90.57             26.6
```

Multiple Regression

```
> attach(swiss)
> swiss_model=lm(Fertility~1)
> add1(swiss_model,.~Agriculture+Examination+Education+Catholic
  + Infant.Mortality)
Single term additions

Model:
Fertility ~ 1
                  Df Sum of Sq    RSS    AIC
<none>                          7178.0 238.34
Agriculture        1     894.8 6283.1 234.09
Examination        1    2994.4 4183.6 214.97
Education          1    3162.7 4015.2 213.04
Catholic           1    1543.3 5634.7 228.97
Infant.Mortality   1    1245.5 5932.4 231.39
```

Multiple Regression

```
> swiss_model=lm(Fertility~Agriculture+Examination+Education+Catholic
  +Infant.Mortality)
> drop1(swiss_model)
Single term deletions

Model:
Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality
                 Df Sum of Sq    RSS    AIC
<none>                       2105.0 190.69
Agriculture       1    307.72 2412.8 195.10
Examination       1     53.03 2158.1 189.86
Education         1   1162.56 3267.6 209.36
Catholic          1    447.71 2552.8 197.75
Infant.Mortality  1    408.75 2513.8 197.03
```

Multiple Regression

```
> step(swiss_model,test="F")
Start:  AIC=190.69
Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality

                  Df Sum of Sq    RSS    AIC F value    Pr(>F)
- Examination      1     53.03 2158.1 189.86  1.0328  0.315462
<none>                       2105.0 190.69
- Agriculture      1    307.72 2412.8 195.10  5.9934  0.018727 *
- Infant.Mortality  1    408.75 2513.8 197.03  7.9612  0.007336 **
- Catholic         1    447.71 2552.8 197.75  8.7200  0.005190 **
- Education        1   1162.56 3267.6 209.36 22.6432 2.431e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=189.86
Fertility ~ Agriculture + Education + Catholic + Infant.Mortality
```

Multiple Regression

```
Step:  AIC=189.86
Fertility ~ Agriculture + Education + Catholic + Infant.Mortality

                  Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                         2158.1 189.86
- Agriculture      1    264.18 2422.2 193.29  5.1413   0.02857 *
- Infant.Mortality 1    409.81 2567.9 196.03  7.9757   0.00722 **
- Catholic         1    956.57 3114.6 205.10 18.6165 9.503e-05 ***
- Education        1   2249.97 4408.0 221.43 43.7886 5.140e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
lm(formula = Fertility ~ Agriculture + Education + Catholic +
    Infant.Mortality)

Coefficients:
     (Intercept)       Agriculture         Education          Catholic
         62.1013           -0.1546           -0.9803            0.1247
Infant.Mortality
          1.0784
```