Simple linear regression

STAT 401A - Statistical Methods for Research Workers

Jarad Niemi

Iowa State University

October 4, 2013

Simple Linear Regression

Recall the One-way ANOVA model:

$$Y_{ij} \stackrel{ind}{\sim} N(\mu_i, \sigma^2)$$

where Y_{ij} is the observation for individual j in group i.

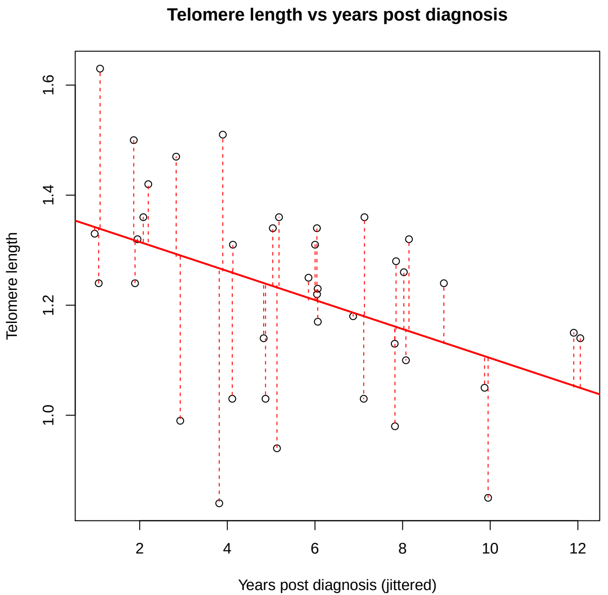
The simple linear regression model is

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

where Y_i and X_i are the response and explanatory variable, respectively, for individual i.

Terminology (all of these are equivalent):

response	explanatory
outcome	covariate
dependent	independent
endogenous	exogenous


Telomere length vs years post diagnosis
Telomere length
1.0
1.2
1.4
1.6
2
4
6
8
10
12
Years post diagnosis (jittered)

Interpretation

$$E[Y_i|X_i = x] = \beta_0 + \beta_1 x \qquad V[Y_i|X_i = x] = \sigma^2$$

- If $X_i = 0$, then $E[Y_i|X_i = 0] = \beta_0$.
  β_0 is the expected response when the explanatory variable is zero.

- If X_i increases from x to $x + 1$, then

$$\begin{array}{rl} E[Y_i|X_i = x+1] & = \beta_0 + \beta_1 x + \beta_1 \\ -\,E[Y_i|X_i = x] & = \beta_0 + \beta_1 x \\ \hline & = \beta_1 \end{array}$$

  β_1 is the expected increase in the response for each unit increase in the explanatory variable.

- σ is the standard deviation of the response for a fixed value of the explanatory variable.

Remove the mean:

$$Y_i = \beta_0 + \beta_1 X_i + e_i \qquad e_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

So

$$e_i = Y_i - (\beta_0 + \beta_1 X_i)$$

which we approximate by the residual

$$r_i = \hat{e}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

The least squares, maximum likelihood, and Bayesian estimators are

$$\begin{aligned} \hat{\beta}_1 &= SXY/SXX \\ \hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1 \overline{X} \\ \hat{\sigma}^2 &= SSE/(n-2) \qquad \text{d.f.} = n-2 \end{aligned}$$

$$\begin{aligned} SXY &= \sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) \\ SXX &= \sum_{i=1}^n (X_i - \overline{X})(X_i - \overline{X}) = \sum_{i=1}^n (X_i - \overline{X})^2 \\ SSE &= \sum_{i=1}^n r_i^2 \end{aligned}$$

$$\begin{aligned} \overline{X} &= \frac{1}{n} \sum_{i=1}^n X_i \\ \overline{Y} &= \frac{1}{n} \sum_{i=1}^n Y_i \end{aligned}$$

How certain are we about $\hat{\beta}_0$ and $\hat{\beta}_1$ being equal to β_0 and β_1?

We quantify this uncertainty using their standard errors:

$$
\begin{aligned}
SE(\beta_0) &= \hat{\sigma}\sqrt{\tfrac{1}{n} + \tfrac{\overline{X}^2}{(n-1)s_X^2}} & \qquad d.f. = n-2 \\
SE(\beta_1) &= \hat{\sigma}\sqrt{\tfrac{1}{(n-1)s_X^2}} & \qquad d.f. = n-2
\end{aligned}
$$

$$
\begin{aligned}
s_X^2 &= SXX/(n-1) \\
s_Y^2 &= SYY/(n-1) \\
SYY &= \textstyle\sum_{i=1}^n (Y_i - \overline{Y})^2
\end{aligned}
$$

$$
\begin{aligned}
r_{XY} &= \tfrac{SXY/(n-1)}{s_X s_Y} & & \qquad \text{correlation coefficient} \\
R^2 &= r_{XY}^2 & = \tfrac{SST-SSE}{SST} & \qquad \text{coefficient of determination} \\
SST &= SYY = \textstyle\sum_{i=1}^n (Y_i - \overline{Y})^2
\end{aligned}
$$

The coefficient of determination is the percentage of the total response variation explained by the explanatory variable(s).

Pvalues and confidence interval

We can compute two-sided pvalues via

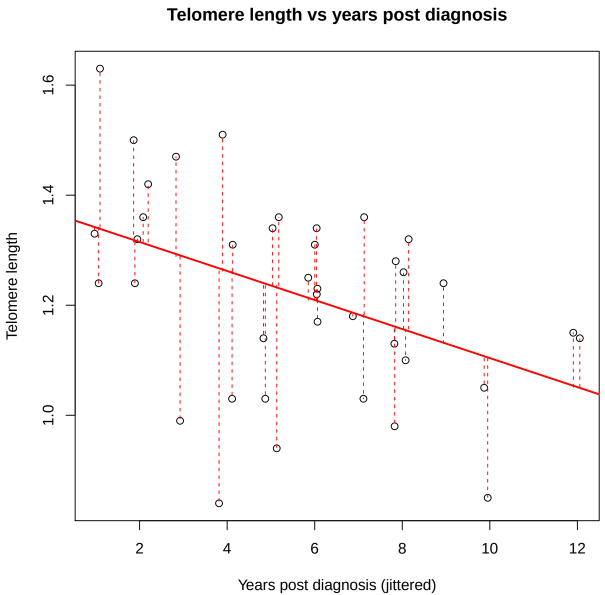
$$2P\left(t_{n-2} > \left|\frac{\hat{\beta}_0}{SE(\beta_0)}\right|\right) \qquad \text{and} \qquad 2P\left(t_{n-2} > \left|\frac{\hat{\beta}_1}{SE(\beta_1)}\right|\right)$$

These test the null hypothesis that the corresponding parameter is zero.

We can construct $100(1-\alpha)\%$ confidence intervals via

$$\hat{\beta}_0 \pm t_{n-2}(1-\alpha/2)SE(\beta_0) \qquad \text{and} \qquad \hat{\beta}_1 \pm t_{n-2}(1-\alpha/2)SE(\beta_1)$$

These provide ranges of the parameter consistent with the data.


Telomere length vs years post diagnosis
Telomere length
1.0
1.2
1.4
1.6
2
4
6
8
10
12
Years post diagnosis (jittered)

```
DATA t;
  INFILE 'telomeres.csv' DSD FIRSTOBS=2;
  INPUT years length;

PROC REG DATA=t;
  MODEL length = years;
  RUN;
```

```
                               The REG Procedure
                                Model: MODEL1
                          Dependent Variable: length

                    Number of Observations Read          39
                    Number of Observations Used          39

                             Analysis of Variance

                                    Sum of           Mean
        Source              DF     Squares         Square    F Value    Pr > F

        Model                1     0.22777        0.22777       8.42    0.0062
        Error               37     1.00033        0.02704
        Corrected Total     38     1.22810

                   Root MSE              0.16443    R-Square     0.1855
                   Dependent Mean        1.22026    Adj R-Sq     0.1634
                   Coeff Var            13.47473


                             Parameter Estimates
                   Parameter     Standard
  Variable    DF    Estimate        Error    t Value    Pr > |t|     95% Confidence Limits

  Intercept    1     1.36768      0.05721      23.91      <.0001      1.25176       1.48360
  years        1    -0.02637      0.00909      -2.90      0.0062     -0.04479      -0.00796
```