Set R01 - Simple linear regression

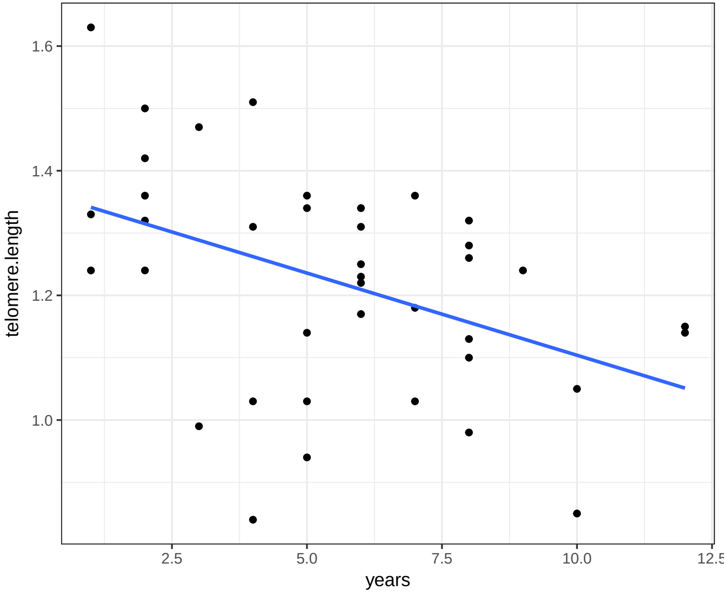
STAT 401 (Engineering) - Iowa State University

March 8, 2017

Telomere length

http://www.pnas.org/content/101/49/17312

> *People who are stressed over long periods tend to look haggard, and it is commonly thought that psychological stress leads to premature aging and the earlier onset of diseases of aging.*
>
> *...*
>
> *This design allowed us to examine the importance of perceived stress and measures of objective stress (caregiving status and chronicity of caregiving stress based on the number of years since a child's diagnosis).*
>
> *...*
>
> *Telomere length values were measured from DNA by a quantitative PCR assay that determines the relative ratio of telomere repeat copy number to single-copy gene copy number (T/S ratio) in experimental samples as compared with a reference DNA sample.*


1.6
1.4
1.2
1.0
telomere.length
2.5
5.0
7.5
10.0
12.5
years

Simple Linear Regression

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

Parameter interpretation

Recall:

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

- 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 the error is

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

which we approximate by the residual

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

The least squares (minimize $\sum_{i=1}^n r_i^2$), maximum likelihood, and Bayesian estimators are

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

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

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

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

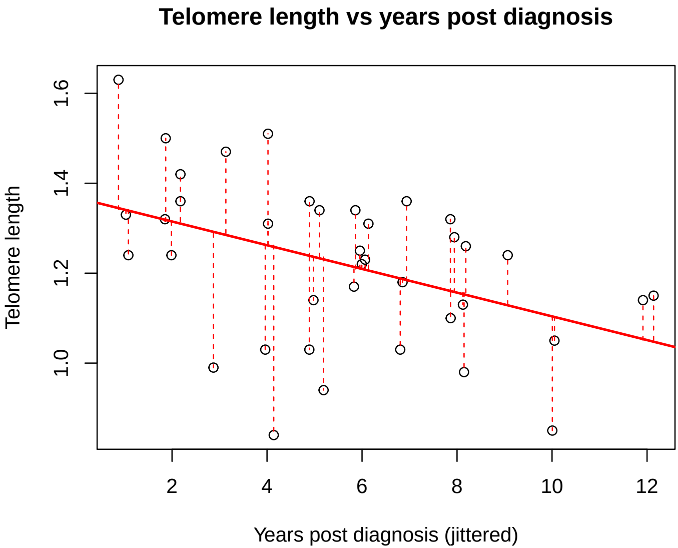
We quantify this uncertainty using their standard errors and posterior standard deviations:

$$\begin{array}{rll}
SE(\beta_0) & = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{(n-1)s_X^2}} & df = n-2 \\
SE(\beta_1) & = \hat{\sigma}\sqrt{\frac{1}{(n-1)s_X^2}} & df = n-2
\end{array}$$

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

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

The coefficient of determination (R^2) is the proportion of the total response variation explained by the explanatory variable(s).


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)

Pvalues and confidence interval

We can compute two-sided pvalues, e.g. $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$, 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 $100(1-a)\%$ two-sided confidence/credible intervals via

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

Calculations by hand in "R"

```
sm <- Telomeres %>% summarize(n = n(), Xbar = mean(years), Ybar = mean(telomere.length),
                              s_X = sd(years), s_Y = sd(telomere.length), r_XY = cor(telomere.length,years))
sm


   n     Xbar     Ybar      s_X       s_Y       r_XY
1 39 5.589744 1.220256 2.935427 0.1797731 -0.4306534
```

$$
\begin{aligned}
SXX &= (n-1)s_x^2 = (39-1)\times 2.9354274^2 = 327.4358974 \\
SYY &= (n-1)s_Y^2 = (39-1)\times 0.1797731^2 = 1.2280974 \\
SXY &= (n-1)s_X s_Y r_{XY} = (39-1)\times 2.9354274 \times 0.1797731 \times -0.4306534 = -8.6358974 \\
\hat\beta_1 &= SXY/SXX = -8.635897/327.4358 = -0.02637432 \\
\hat\beta_0 &= \overline{Y} - \hat\beta_1 \overline{X} = 1.220256 - (-0.02637432)\times 5.589744 = 1.367682 \\
R^2 &= r_{XY}^2 = (-0.4306534)^2 = 0.1854624 \\
SSE &= SYY(1-R^2) = 1.228098(1-0.1854624) = 1.000332 \\
\hat\sigma^2 &= SSE/(n-2) = 1.000332/(39-2) = 0.027036 \\
\hat\sigma &= \sqrt{\hat\sigma^2} = \sqrt{0.027036} = 0.1644263 \\
SE(\hat\beta_0) &= \hat\sigma\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{(n-1)s_x^2}} = 0.1644263\sqrt{\frac{1}{39} + \frac{5.589744^2}{327.4358}} = 0.05721115 \\
SE(\hat\beta_1) &= \hat\sigma\sqrt{\frac{1}{(n-1)s_x^2}} = 0.1644263\sqrt{\frac{1}{327.4358}} = 0.009086742 \\
p_{H_0:\beta_0=0} &= 2P\left(t_{n-2} < -\left|\frac{1.367682}{0.05721115}\right|\right) = 2P(t_{37} < -23.90586) < 0.0001 \\
p_{H_0:\beta_1=0} &= 2P\left(t_{n-2} < -\left|\frac{-0.02637432}{0.009086742}\right|\right) = 2P(t_{37} < -2.902506) < 0.0062 \\
CI_{95\%\ \beta_0} &= \hat\beta_0 \pm t_{n-2,1-a/2} SE(\hat\beta_0) = 1.367682 \pm 2.026192 \times 0.05721115 = (1.251761, 1.483603) \\
CI_{95\%\ \beta_1} &= \hat\beta_1 \pm t_{n-2,1-a/2} SE(\hat\beta_1) = -0.02637432 \pm 2.026192 \times 0.009086742 = (-0.044785804 - 0.007
\end{aligned}
$$

Regression in R

```
m = lm(telomere.length~years, Telomeres)
summary(m)


Call:
lm(formula = telomere.length ~ years, data = Telomeres)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42218 -0.08537  0.02056  0.10738  0.28869

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.367682   0.057211  23.906   <2e-16 ***
years       -0.026374   0.009087  -2.903   0.0062 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1644 on 37 degrees of freedom
Multiple R-squared:  0.1855, Adjusted R-squared:  0.1634
F-statistic: 8.425 on 1 and 37 DF,  p-value: 0.006205


confint(m)


                  2.5 %       97.5 %
(Intercept)  1.25176134  1.483602799
years       -0.04478579 -0.007962836
```

Conclusion

Telomere length at the time of diagnosis of a child's chronic illness is estimated to be 1.37 with a 95% confidence interval of (1.25, 1.48). For each year since diagnosis, the telomere length decreases by 0.026 with a 95% confidence interval of (0.008, 0.045) on average. The proportional of variability in telomere length described by years since diagnosis is 18.5%.

http://www.pnas.org/content/101/49/17312

> *The zero-order correlation between chronicity of caregiving [years] and mean telomere length, r,is -0.445 (P <0.01). [$R^2 = 0.198$ was shown in the plot.]*

Remark I'm guessing our analysis and that reported in the paper don't match exactly due to a discrepancy in the data.

Summary

- 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.

- Know how to use R to obtain $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$, R^2, pvalues, CIs, etc.

- Interpret R output
  - At a value of zero for the explanatory variable $(X_i = 0)$, β_0 is the expected value for the response (Y_i).
  - For each unit increase in the explanatory variable value, β_1 is the expected increase in the response.
  - At a constant value of the explanatory variable, σ^2 is the variance of the responses.
  - The coefficient of determination (R^2) is the percentage of the total response variation explained by the explanatory variable(s).

What is $E[Y|X=x]$?

We know $\beta_0 = E[Y|X=0]$, but what about $X = x$?

$$E[Y|X=x] = \beta_0 + \beta_1 x$$

which we can estimate via

$$E[\widehat{Y|X=}\,x] = \hat{\beta}_0 + \hat{\beta}_1 x$$

but there is uncertainty in both β_0 and β_1. So the standard error of $E[Y|X=x]$ is

$$SE(E[Y|X=x]) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(\overline{X}-x)^2}{(n-1)s_X^2}}$$

and a $100(1-a)$% confidence interval is

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{n-2,1-a/2} SE(E[Y|X=x])$$

What do we predict about Y at $X = x$?

On the last slide, we calculated $E[Y|X = x]$ and it's uncertainty, but if we are trying to predict a new observation, we need to account for the sampling variablity σ^2. Thus a prediction about Y at a new $X = x$ is still

$$Pred\{Y|X = x\} = \hat{\beta}_0 + \hat{\beta}_1 x$$

but the uncertainty includes the variability due to σ^2. So the standard error of $Pred\{Y|X = x\}$ is

$$SE(Pred\{Y|X = x\}) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(\overline{X} - x)^2}{(n-1)s_X^2}}$$

and a $100(1 - a)\%$ confidence interval is

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{n-2,1-a/2} SE(Pred\{Y|X = x\}).$$

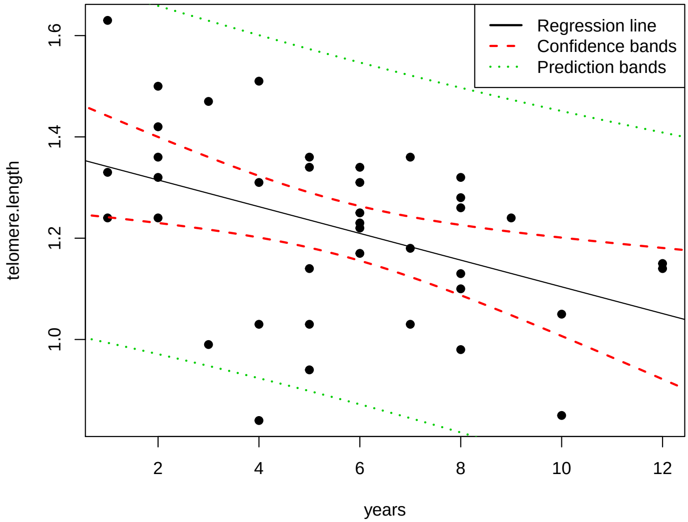
Confidence and prediction intervals for r

```
m = lm(telomere.length~years, Telomeres)
new = data.frame(years=4)
new %>% bind_cols(predict(m, new, interval="confidence") %>% as.data.frame)
```

```
  years      fit      lwr      upr
1     4 1.262185 1.201335 1.323035
```

```
new %>% bind_cols(predict(m, new, interval="prediction") %>% as.data.frame)
```

```
  years      fit       lwr      upr
1     4 1.262185 0.9235142 1.600855
```


Regression line
Confidence bands
Prediction bands
telomere.length
1.0
1.2
1.4
1.6
2
4
6
8
10
12
years

Shifting the intercept

The intercept (β_0) is the expected response when the explanatory variable is zero.

So, if we change our explanatory variable, we change the interpretation of our intercept, e.g. if, instead of using number of years since diagnosis, we use "number of years since diagnosis minus 4", then our intercept is the expected response at 4 years since diagnosis.

Let x be number of years since diagnosis, then

$$E[Y|X = x] = \tilde{\beta}_0 + \tilde{\beta}_1(x - 4) = (\beta_0 - 4\beta_1) + \beta_1 x$$

so our new parameters for the mean are

- intercept $\tilde{\beta}_0 = (\beta_0 - 4\beta_1)$and
- slope $\tilde{\beta}_1 = \beta_1$ (unchanged).

```
m0 = lm(telomere.length ~   years   , Telomeres)
m4 = lm(telomere.length ~ I(years-4), Telomeres)

coef(m0)


(Intercept)       years
 1.36768207 -0.02637431


coef(m4)


 (Intercept) I(years - 4)
  1.26218481  -0.02637431


confint(m0)


                  2.5 %       97.5 %
(Intercept)  1.25176134  1.483602799
years       -0.04478579 -0.007962836


confint(m4)


                   2.5 %       97.5 %
(Intercept)   1.20133473  1.323034890
I(years - 4) -0.04478579 -0.007962836
```