

R01 - Simple linear regression

STAT 5870 (Engineering)
Iowa State University

October 29, 2024

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 [as measured by decreased telomere length]

...

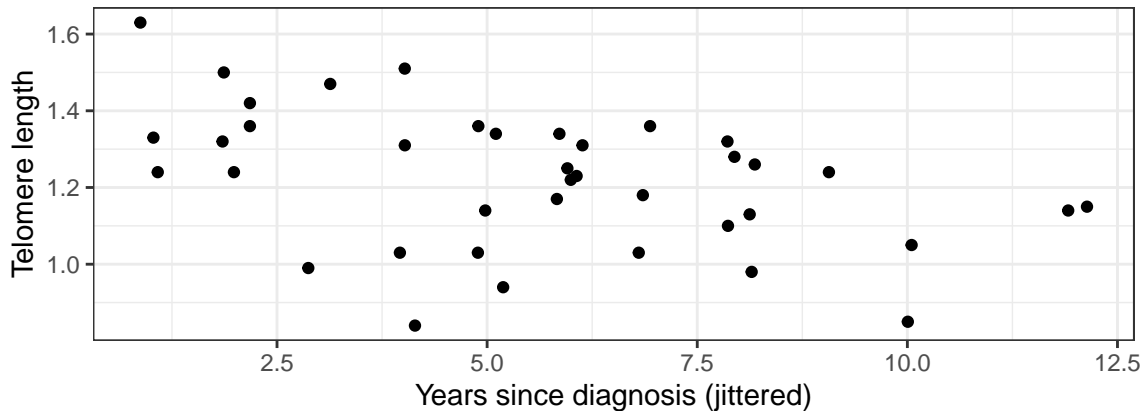
examine the importance of ... caregiving stress (...number of years since a child's diagnosis [of a chronic disease]) [on telomere length]

...

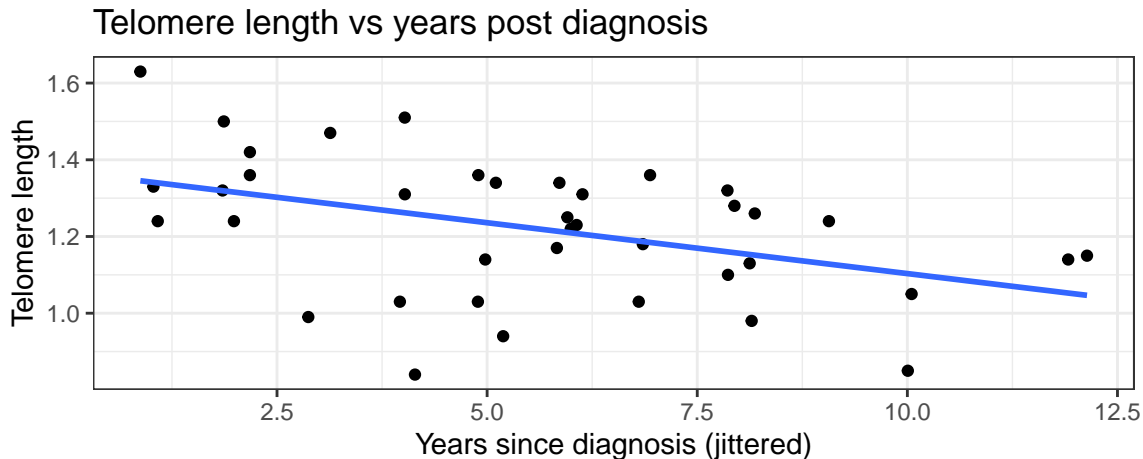
Telomere length values were measured from DNA by a quantitative PCR assay that determines the relative ratio of telomere repeat copy number to single-copy gene copy number (T/S ratio) in experimental samples as compared with a reference DNA sample.

Data

Telomere length vs years post diagnosis



Data with regression line



Simple Linear Regression

The **simple linear regression** model is

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

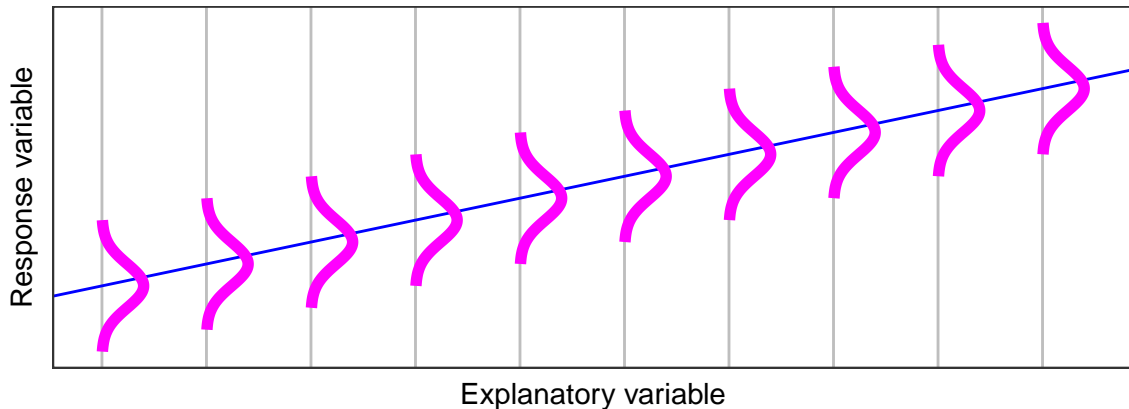
where Y_i and X_i are the response and explanatory variable, respectively, for individual i .

Terminology (all of these are equivalent):

response	explanatory
outcome	covariate
dependent	independent
endogenous	exogenous

Simple linear regression - visualized

Simple linear regression model



Parameter interpretation

Recall:

$$E[Y_i|X_i = x] = \beta_0 + \beta_1 x \quad \text{Var}[Y_i|X_i = x] = \sigma^2$$

- 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

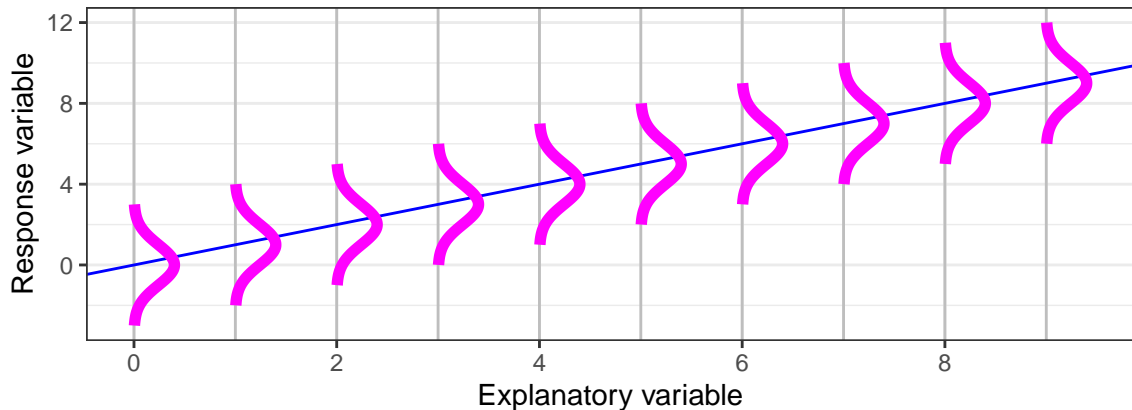
$$\frac{\begin{array}{l} E[Y_i|X_i = x + 1] \\ - E[Y_i|X_i = x] \end{array}}{= \quad \quad \quad} = \frac{\beta_0 + \beta_1 x + \beta_1}{\beta_0 + \beta_1 x} = \beta_1$$

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.

Simple linear regression - visualized

Simple linear regression model



Errors v residuals

Remove the mean:

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

So the error is

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

which we approximate by the **residual**

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

These residuals we will use for a number of purposes including

- assessing model assumptions,
- identifying outliers, and
- estimating error variance.

Estimators

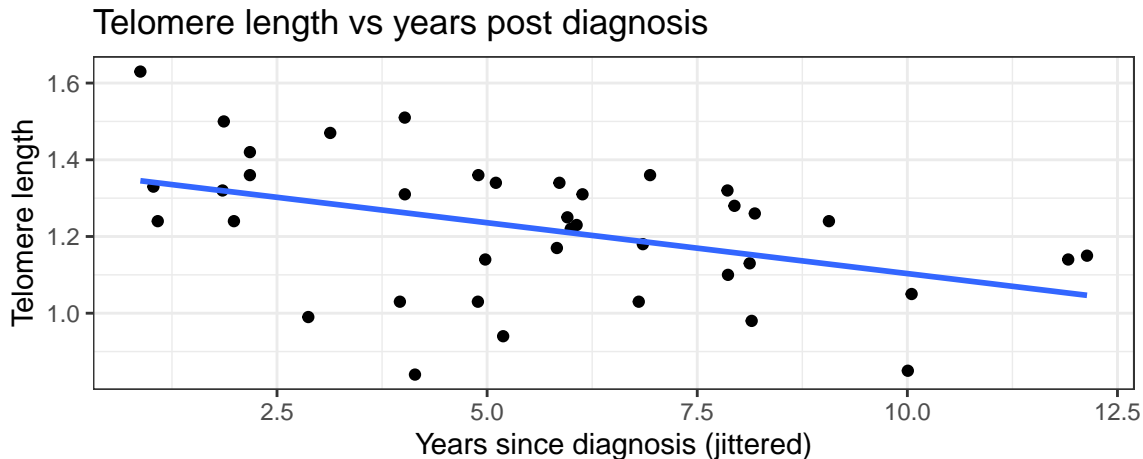
The least squares (minimize $\sum_{i=1}^n r_i^2$), maximum likelihood, and Bayesian estimators (prior $1/\sigma^2$) are

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

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

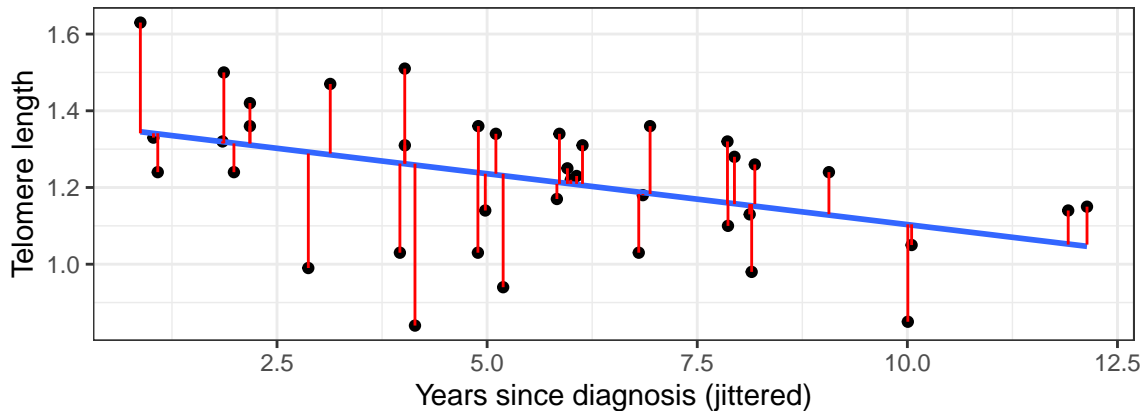
$$\begin{aligned}SXY &= \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \\ SXX &= \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X}) = \sum_{i=1}^n (X_i - \bar{X})^2 \\ SSE &= \sum_{i=1}^n r_i^2\end{aligned}$$

Residuals



Residuals

Telomere length vs years post diagnosis



How certain are we about $\hat{\beta}_0$ and $\hat{\beta}_1$?

We quantify this uncertainty using their standard errors (or posterior scale parameters):

$$SE(\hat{\beta}_0) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}} \quad df = n - 2$$

$$SE(\hat{\beta}_1) = \hat{\sigma} \sqrt{\frac{1}{(n-1)s_X^2}} \quad df = n - 2$$

$$s_X^2 = SXX/(n-1)$$

$$s_Y^2 = SYX/(n-1)$$

$$SYY = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$r_{XY} = \frac{SXY/(n-1)}{s_X s_Y} \quad \text{correlation coefficient}$$

$$R^2 = r_{XY}^2 = \frac{SST - SSE}{SST} \quad \text{coefficient of determination}$$

$$SST = SYX = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

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

Default Bayesian analysis of the simple linear regression model

If we assume the default prior $p(\beta_0, \beta_1, \sigma^2) \propto 1/\sigma^2$, then the marginal posteriors for the mean parameters are

$$\beta_j|y \sim t_{n-2}(\hat{\beta}_j, SE(\hat{\beta}_j)^2).$$

We can construct a $100(1-a)\%$ two-sided credible interval for β_j via

$$\hat{\beta}_j \pm t_{n-2, 1-a/2} SE(\hat{\beta}_j)$$

where $P(T_{n-2} < t_{n-2, 1-a/2}) = 1 - a/2$ for $T_{n-2} \sim t_{n-2}$.

We can compute posterior probabilities via

$$P(\beta_j > b_j|y) = P\left(T_{n-2} > \frac{b_j - \hat{\beta}_j}{SE(\hat{\beta}_j)}\right) \quad \text{or} \quad P(\beta_j < b_j|y) = P\left(T_{n-2} < \frac{b_j - \hat{\beta}_j}{SE(\hat{\beta}_j)}\right)$$

often $b_j = 0$.

p-values and confidence interval

We can construct a $100(1 - \alpha)\%$ two-sided confidence interval for β_j via

$$\hat{\beta}_j \pm t_{n-2, 1-\alpha/2} SE(\hat{\beta}_j).$$

We can compute one-sided p-values,

$H_0 : \beta_j \geq b_j$ vs $H_A : \beta_j < b_j$ has

and $H_0 : \beta_j \leq b_j$ vs $H_A : \beta_j > b_j$ has

$$p\text{-value} = P\left(T_{n-2} < \frac{\hat{\beta}_j - b_j}{SE(\hat{\beta}_j)}\right)$$

$$p\text{-value} = P\left(T_{n-2} > \frac{\hat{\beta}_j - b_j}{SE(\hat{\beta}_j)}\right)$$

or two-sided p-values for $H_0 : \beta_j = b_j$ vs $H_A : \beta_j \neq b_j$:

$$= 2 \times \min \left\{ P\left(T_{n-2} > \frac{\hat{\beta}_j - b_j}{SE(\hat{\beta}_j)}\right), P\left(T_{n-2} < \frac{\hat{\beta}_j - b_j}{SE(\hat{\beta}_j)}\right) \right\} = 2 \times P\left(T_{n-2} < -\left| \frac{\hat{\beta}_j - b_j}{SE(\hat{\beta}_j)} \right|\right)$$

software default is usually $b_j = 0$.

Calculations “by hand” in R

```
n      = nrow(Telomeres)
Xbar    = mean(Telomeres$years)
Ybar    = mean(Telomeres$telomere.length)
s_X     = sd(Telomeres$years)
s_Y     = sd(Telomeres$telomere.length)
r_XY    = cor(Telomeres$telomere.length, Telomeres$years)

SXX = (n-1)*s_X^2
SYY = (n-1)*s_Y^2
SXY = (n-1)*s_X*s_Y*r_XY

beta1 = SXY/SXX
beta0 = Ybar - beta1 * Xbar

R2    = r_XY^2
SSE   = SYY*(1-R2)

sigma2 = SSE/(n-2)
sigma  = sqrt(sigma2)

SE_beta0 = sigma*sqrt(1/n + Xbar^2/((n-1)*s_X^2))
SE_beta1 = sigma*sqrt(1/((n-1)*s_X^2))
```


Calculations “by hand” in R (continued)

```
# 95% CI for beta0  
beta0 + c(-1, 1) * qt(.975, df = n-2) * SE_beta0
```

```
[1] 1.251761 1.483603
```

```
# 95% CI for beta1  
beta1 + c(-1, 1) * qt(.975, df = n-2) * SE_beta1
```

```
[1] -0.044785794 -0.007962836
```

```
# pvalue for H0: beta0 <= 0 and P(beta0 > 0 | y)  
pt(beta0 / SE_beta0, df = n - 2)
```

```
[1] 1
```

```
# pvalue for H0: beta1 <= 0 and P(beta1 > 0 | y)  
pt(beta1 / SE_beta1, df = n - 2)
```

```
[1] 0.003102353
```

```
# pvalue for H1: beta0 = 0  
2 * pt(-abs(beta1 / SE_beta1), df = n - 2)
```

```
[1] 0.006204706
```

Calculations by hand

$$\begin{aligned}
 SXX &= (n-1)s_x^2 = (39-1) \times 2.9354274^2 = 327.4358974 \\
 SY Y &= (n-1)s_Y^2 = (39-1) \times 0.1797731^2 = 1.2280974 \\
 SXY &= (n-1)s_X s_Y r_{XY} = (39-1) \times 2.9354274 \times 0.1797731 \times -0.4306534 = -8.6358974 \\
 \hat{\beta}_1 &= SXY/SXX = -8.6358974/327.4358974 = -0.0263743 \\
 \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} = 1.2202564 - (-0.0263743) \times 5.5897436 = 1.3676821 \\
 R^2 &= r_{XY}^2 = (-0.4306534)^2 = 0.1854624 \\
 SSE &= SY Y (1 - R^2) = 1.2280974(1 - 0.1854624) = 1.0003316 \\
 \hat{\sigma}^2 &= SSE/(n-2) = 1.0003316/(39-2) = 0.027036 \\
 \hat{\sigma} &= \sqrt{\hat{\sigma}^2} = \sqrt{0.027036} = 0.1644262 \\
 SE(\hat{\beta}_0) &= \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2}} = 0.1644262 \sqrt{\frac{1}{39} + \frac{5.5897436^2}{(39-1) \times 2.9354274^2}} = 0.0572111 \\
 SE(\hat{\beta}_1) &= \hat{\sigma} \sqrt{\frac{1}{(n-1)s_x^2}} = 0.1644262 \sqrt{\frac{1}{(39-1) \times 2.9354274^2}} = 0.0090867 \\
 p_{H_A: \beta_0 \neq 0} &= 2P\left(T_{n-2} < -\left|\frac{\hat{\beta}_0}{SE(\hat{\beta}_0)}\right|\right) = 2P(t_{37} < -23.9058799) = 4.2740348 \times 10^{-24} \\
 p_{H_A: \beta_1 \neq 0} &= 2P\left(T_{n-2} < -\left|\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}\right|\right) = 2P(t_{37} < -2.9025065) = 0.0062047 \\
 CI_{95\% \beta_0} &= \hat{\beta}_0 \pm t_{n-2, 1-\alpha/2} SE(\hat{\beta}_0) \\
 &= 1.3676821 \pm 2.0261925 \times 0.0572111 = (1.2517613, 1.4836028) \\
 CI_{95\% \beta_1} &= \hat{\beta}_1 \pm t_{n-2, 1-\alpha/2} SE(\hat{\beta}_1) \\
 &= -0.0263743 \pm 2.0261925 \times 0.0090867 = (-0.0447858, -0.0079628)
 \end{aligned}$$

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

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

The correlation between chronicity of caregiving and mean telomere length 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 , p -values, CIs, etc.
- Interpret regression output:
 - β_0 is the expected value for the response when the explanatory variable is 0.
 - β_1 is the expected increase in the response for each unit increase in the explanatory variable.
 - σ is the standard deviation of responses around their mean.
 - R^2 is the proportion of the total variation of the response variable explained by the model.