

R01 - Simple linear regression:

Choosing explanatory variables

STAT 587 (Engineering)
Iowa State University

October 19, 2020

Simple linear regression

Let

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

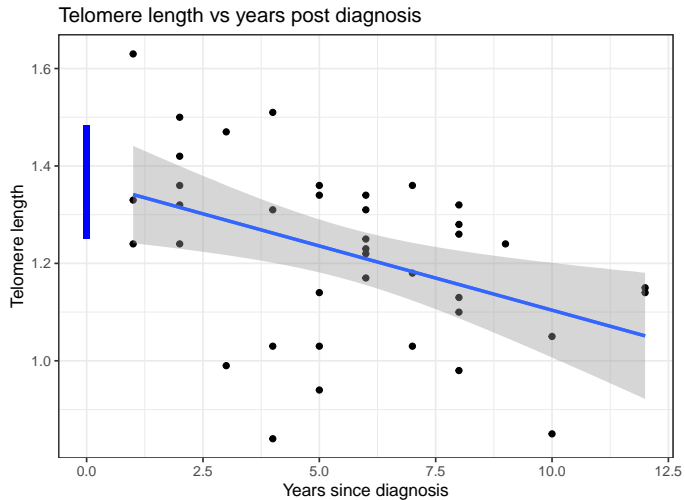
Possible choices for f :

- quadratic: $f(x) = x^2$
- logarithmic: $f(x) = \log(x)$
- centered: $f(x) = x - m$
- scaled: $f(x) = x/s$

Quadratic relationship



Logarithmic relationship



Shifting the intercept

The intercept is the expected response when the explanatory variable is zero. If we use

$$f(x) = x - m,$$

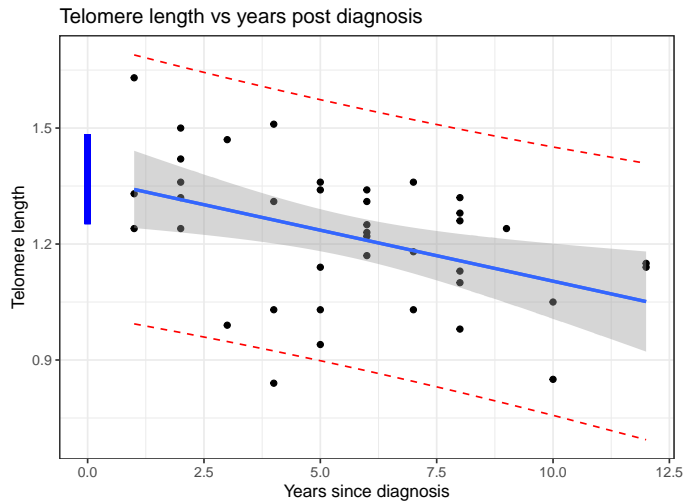
then the new intercept is the expected response when the explanatory variable is m .

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

so our new parameters for the mean are

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

Telomere data



Telomere data: shifting the intercept

```
m0 = lm(telomere.length ~ years, abd::Telomeres)
m4 = lm(telomere.length ~ I(years-5), abd::Telomeres)
```

```
coef(m0)
```

```
(Intercept)      years
 1.36768207 -0.02637431
```

```
coef(m4)
```

```
(Intercept) I(years - 5)
 1.23581049 -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)  1.18136856  1.290252429
I(years - 5) -0.04478579 -0.007962836
```

Rescaling the slope

The slope is the expected increase in the response when the explanatory variable increases by 1. If we use

$$f(x) = x/s,$$

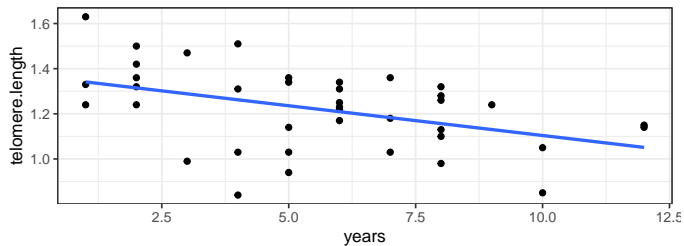
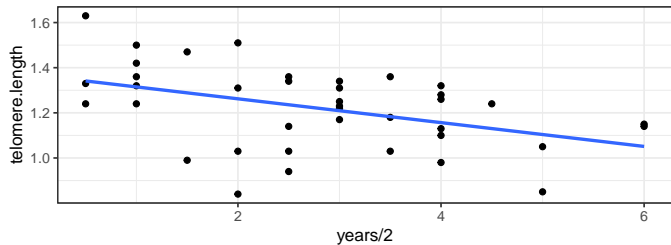
then the new slope is the expected increase in the response when the explanatory variable increases by s .

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

so our new parameters are

- intercept $\tilde{\beta}_0 = \beta_0$ (unchanged) but
- slope $\tilde{\beta}_1 = \beta_1/s$.

Telomere data: rescaling the slope



Telomere data: rescaling the slope

```
m0 = lm(telomere.length ~ years, abd::Telomeres)
m4 = lm(telomere.length ~ I(years/2), abd::Telomeres)
```

```
coef(m0)
```

```
(Intercept)      years
 1.36768207 -0.02637431
```

```
coef(m4)
```

```
(Intercept)  I(years/2)
 1.36768207 -0.05274863
```

```
confint(m0)
```

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

```
confint(m4)
```

```
                2.5 %      97.5 %
(Intercept)  1.25176134  1.48360280
I(years/2)   -0.08957159 -0.01592567
```

Summary

Let

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

Choose f based on

- Scientific understanding
- Interpretability
- Diagnostics