Homework 3 solutions

STAT3530 Applied Linear Models

Solutions are shown in blue.

Fruit flies

1. The `flies` data give `thorax` length in millimeters and `longevity` in days for 124 male fruit flies. This problem considers the hypothesis that larger flies live longer on average.

   a. Write the linear model that facilitates testing the hypothesis of interest.

$$\text{longevity}_i = \beta_0 + \beta_1 \,\text{thorax}_i + \epsilon_i, \qquad \mathbb{E}[\epsilon] = 0, \qquad \text{Var}[\epsilon] = \sigma^2 I$$

   The hypothesis that larger flies live longer is a claim about β_1: larger flies live longer on average if $\beta_1 > 0$.

   b. Fit the model and report the estimates and standard errors.

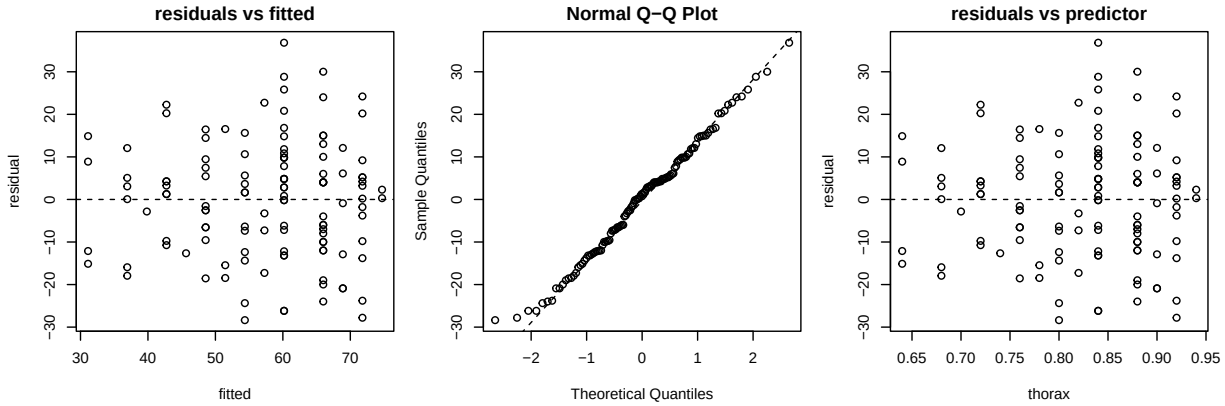
```
fit <- lm(longevity ~ thorax, data = flies)
summary(fit)$coefficients
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -61.85769   13.37405 -4.625202 9.391576e-06
thorax      145.27706   16.19336  8.971397 4.274723e-15
```

   $\hat{\beta}_0 = -61.86$ (SE 13.37) and $\hat{\beta}_1 = 145.28$ days per mm (SE 16.19), with $\hat{\sigma} = 13.65$ on 122 degrees of freedom. A millimeter is an enormous change here — the data span 0.64 to 0.94 mm — which is why part (e) rescales it.

c. Check the residual diagnostics and comment on model assumptions.

```
par(mfrow = c(1, 3), mar = c(4, 4, 2, 1))
plot(fitted(fit), resid(fit), xlab = 'fitted', ylab = 'residual',
     main = 'residuals vs fitted')
abline(h = 0, lty = 2)
qqnorm(resid(fit)); qqline(resid(fit), lty = 2)
plot(flies$thorax, resid(fit), xlab = 'thorax', ylab = 'residual',
     main = 'residuals vs predictor')
abline(h = 0, lty = 2)
```



A shapeless band of even width and a QQ plot close to the line: linearity, constant variance, and normality all look fine. The largest standardized residual is 2.71 and the largest Cook's distance 0.05, so no fly drives the fit. The model is well specified and the inference below can be taken at face value.

d. Test the hypothesis that larger flies live longer on average. State the hypotheses, report the test statistic and p-value, and interpret the result.

$H_0 : \beta_1 = 0$ against $H_1 : \beta_1 > 0$; the alternative is one-sided because the claim is that larger flies live *longer*.

```
summary(fit)$coefficients['thorax', ]
```

```
    Estimate   Std. Error      t value     Pr(>|t|)
1.452771e+02 1.619336e+01 8.971397e+00 4.274723e-15
```

```
summary(fit)$coefficients['thorax', 'Pr(>|t|)'] / 2   # one-sided
```

```
[1] 2.137361e-15
```

$t = 8.97$ on 122 degrees of freedom, and halving the two-sided p-value `summary()` reports gives $p \approx 2 \times 10^{-15}$. Overwhelming evidence that larger flies live longer on average.

e. Report and interpret a 95% confidence interval for the change in mean longevity per 0.1mm increase in thorax length.

Scale the interval for β_1 by 0.1:

```
0.1 * confint(fit)['thorax', ]
```

```
   2.5 %   97.5 %
11.32207 17.73334
```

Each additional 0.1 mm of thorax is associated with between 11.3 and 17.7 more days of mean longevity, with a point estimate of 14.5 days.

f. Compute and interpret a 95% interval for the expected lifespan of a fly with a 0.8mm thorax.

"Expected" means a mean, so this is a confidence interval, not a prediction interval:

```
predict(fit, data.frame(thorax = 0.8), interval = 'confidence')
```

```
       fit      lwr      upr
1 54.36395 51.83262 56.89528
```

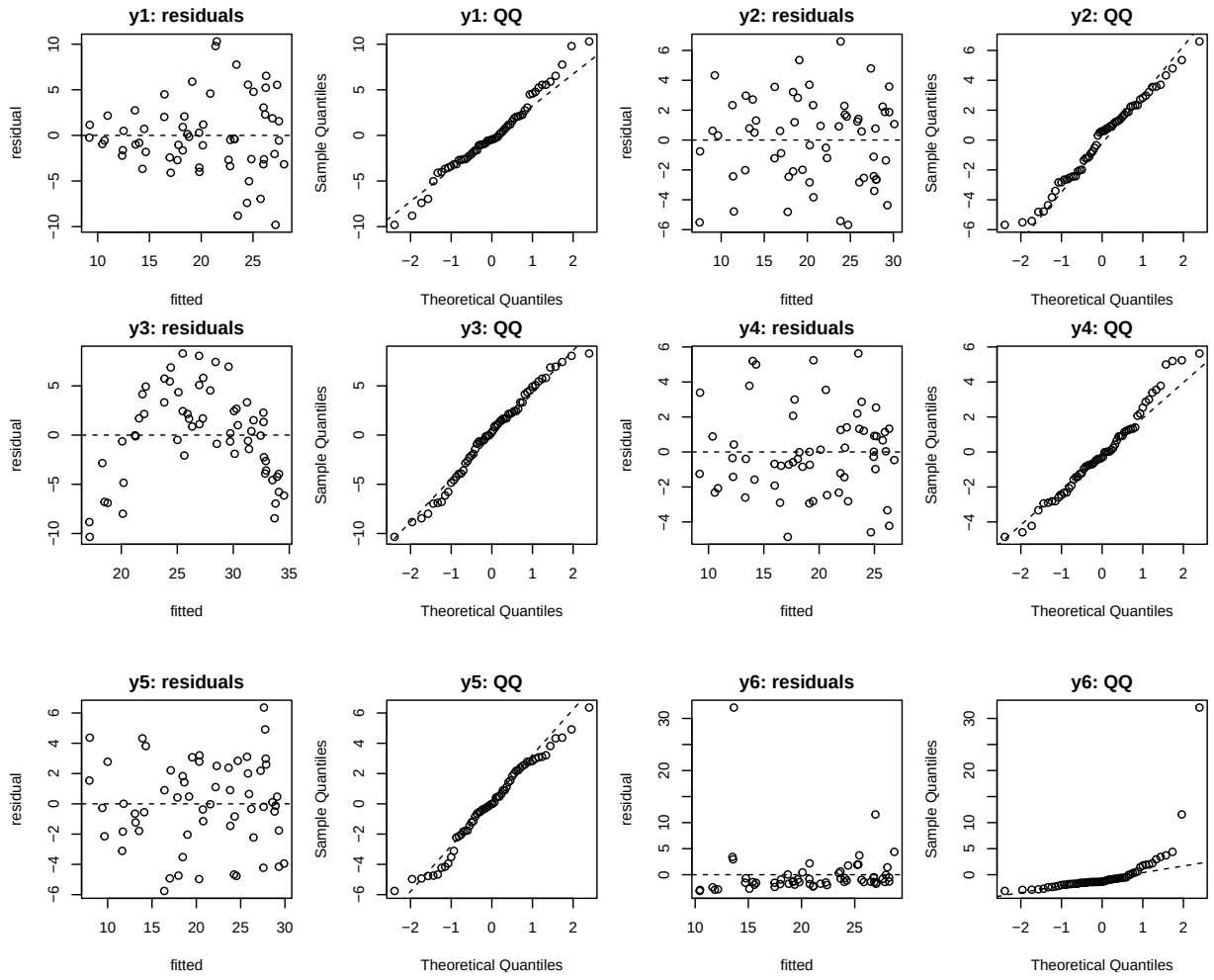
Mean longevity among flies with a 0.8 mm thorax is estimated at 54.4 days, with a 95% interval of $(51.8, 56.9)$. For a *single* fly the prediction interval would be $(27.2, 81.5)$ — far wider, since it also carries the scatter of flies about the line.

Spot the issue

2. For each response, fit the simple linear model and check the assumptions.

   a. Make residual diagnostic plots for each of the six fits.

```
par(mfrow = c(2, 4), mar = c(4, 4, 2, 1), cex = 0.7)
for (k in 1:6) {
  f <- lm(sim[[paste0('y', k)]] ~ sim$x)
  plot(fitted(f), resid(f), xlab = 'fitted', ylab = 'residual',
       main = paste0('y', k, ': residuals'))
  abline(h = 0, lty = 2)
  qqnorm(resid(f), main = paste0('y', k, ': QQ'))
  qqline(resid(f), lty = 2)
}
```



b. Complete the table. Write "none" when nothing is wrong.

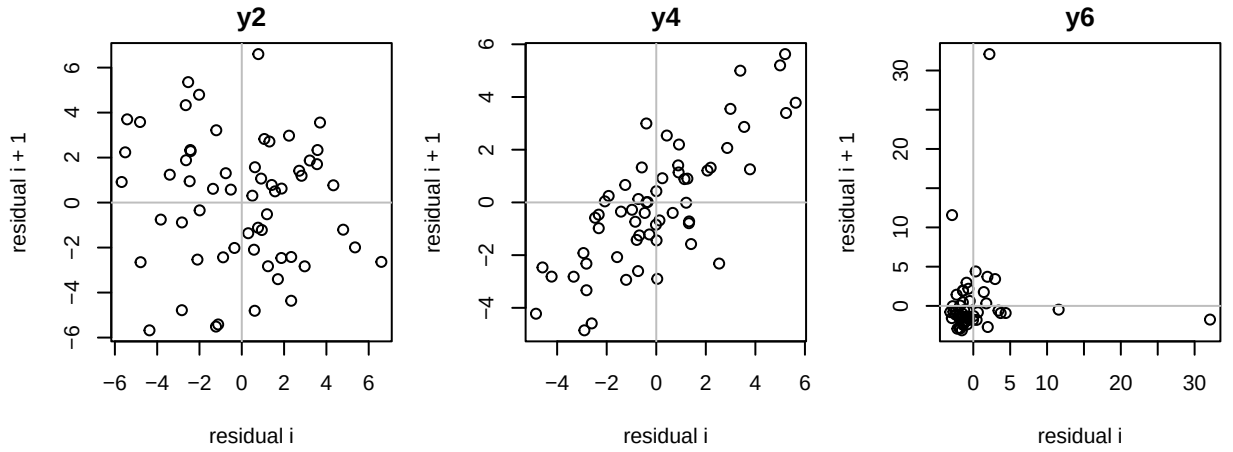
response	assumption that fails	which plot showed it	what you saw
y1	constant variance	residuals vs fitted	spread widens steadily from left to right
y2	none	—	shapeless band, QQ on the line
y3	linearity	residuals vs fitted	a clear arc: negative at both ends, positive in the middle
y4	independence	lag plot of the residuals	neighboring residuals track each other
y5	none	—	shapeless band, QQ on the line
y6	normality	normal QQ	one residual near +32 and a badly skewed right tail

y4 looks fine in both plots above; the problem only appears when residuals are compared with their neighbors, which is part (c). y6's residual plot also looks uneven, but that is the one huge residual distorting the picture, not a variance pattern.

c. Does the independence assumption fail for any of the models? If so, how did you detect it?

Yes, for y4. Each residual plotted against the one before it:

```
par(mfrow = c(1, 3), mar = c(4, 4, 2, 1), cex = 0.7)
for (k in c(2, 4, 6)) {
  e <- resid(lm(sim[[paste0('y', k)]] ~ sim$x))
  plot(head(e, -1), tail(e, -1), xlab = 'residual i', ylab = 'residual i + 1',
       main = paste0('y', k))
  abline(h = 0, v = 0, col = 'grey')
}
```




```
e4 <- resid(lm(y4 ~ x, data = sim))
cor(head(e4, -1), tail(e4, -1))
```

```
[1] 0.7819974
```

The lag-1 correlation is 0.78 for `y4` against -0.11 and 0.04 for the others. The estimates stay unbiased, but the reported standard errors are too small and any test or interval built from them is unreliable.

Transformations

Kleiber's law

3. The `kleiber` data give body mass (kg) and resting metabolic rate (kJ/day) for 95 animal species.

   a. Report the correlation between metabolism and mass on the original scale. Does this suggest a linear relationship?

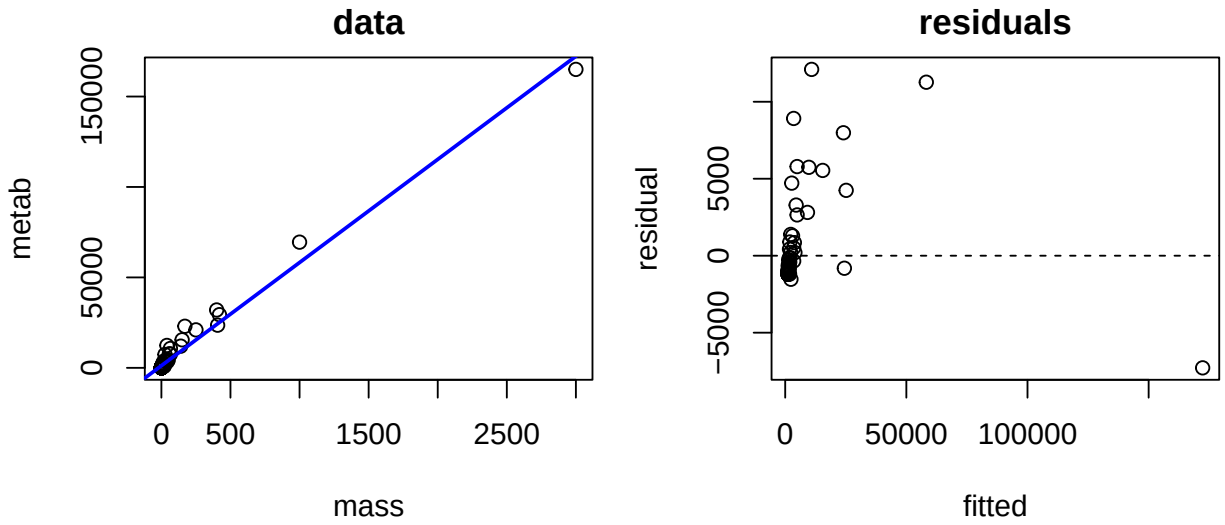
```
cor(kleiber$metab, kleiber$mass)
```

```
[1] 0.989362
```

The correlation is 0.989, which does not establish linearity. With masses spanning six orders of magnitude, a few huge species at one end force a correlation near 1 whatever the shape in between. Linearity is settled by a plot, not a correlation.

   b. Now fit the model `metab ~ mass` and make the residual-versus-fitted plot. Does this suggest a linear relationship? Explain.

```
raw <- lm(metab ~ mass, data = kleiber)
par(mfrow = c(1, 2), mar = c(4, 4, 2, 1))
plot(metab ~ mass, data = kleiber, main = 'data')
abline(raw, col = 'blue', lwd = 2)
plot(fitted(raw), resid(raw), xlab = 'fitted', ylab = 'residual',
     main = 'residuals')
abline(h = 0, lty = 2)
```



No. Almost every species is crushed into the bottom left corner, and the residuals bend away from zero while fanning out. The straight-line model is badly misspecified despite that correlation of 0.989.

c. Write the model implied by the theory as a linear model and identify the parameters corresponding to β_1 and c.

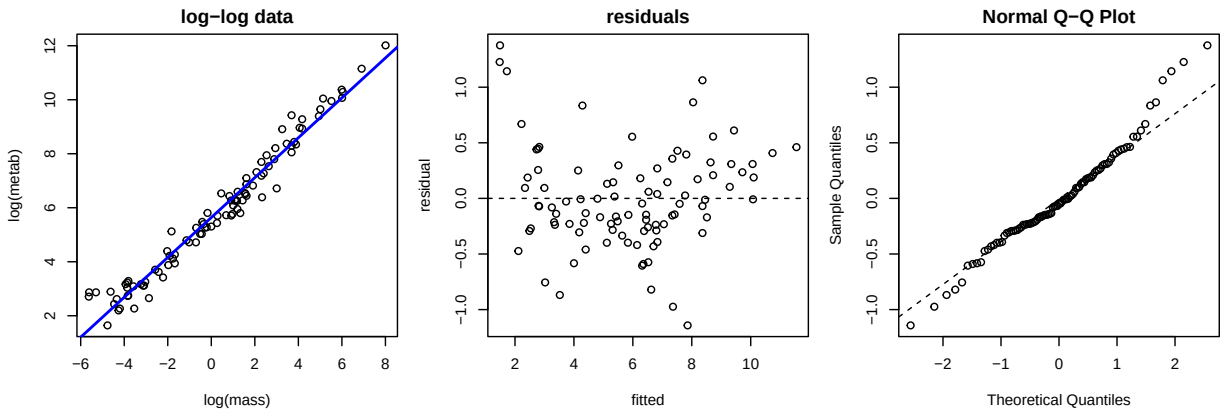
Take logs of metab $= c \cdot \text{mass}^{\beta_1}$:

$$\log(\text{metab}) = \log(c) + \beta_1 \log(\text{mass})$$

a simple linear model in log(mass), with intercept $\beta_0 = \log(c)$, so $c = e^{\beta_0}$. The disputed exponent is the slope β_1.

d. Fit the model in (b) and check the residual diagnostics. Does the model seem sufficiently well-specified for inference on β_1? Why or why not?

```
fit.k <- lm(log(metab) ~ log(mass), data = kleiber)
par(mfrow = c(1, 3), mar = c(4, 4, 2, 1))
plot(log(metab) ~ log(mass), data = kleiber, main = 'log-log data')
abline(fit.k, col = 'blue', lwd = 2)
plot(fitted(fit.k), resid(fit.k), xlab = 'fitted', ylab = 'residual',
     main = 'residuals')
abline(h = 0, lty = 2)
qqnorm(resid(fit.k)); qqline(resid(fit.k), lty = 2)
```



On the log scale the points fall along a straight line across the whole range of masses, the residuals form a band of roughly even width instead of the fan in (b), and the QQ plot is close to the line with slightly heavy tails.

A shallow bow remains in the residuals, driven mostly by the smallest mammals, so the fit is not perfect. It is small relative to the scatter and no observation dominates, so the model is good enough for inference on β_1.

e. Report and interpret a 95% confidence interval for β_1. Which (if either) theory does this support?

```
summary(fit.k)$coefficients
```

```
             Estimate Std. Error   t value      Pr(>|t|)
(Intercept) 5.6383307 0.04709325 119.72693 1.121821e-103
log(mass)   0.7387436 0.01461957  50.53116  2.018934e-69
```

```
confint(fit.k)['log(mass)', ]
```

```
    2.5 %    97.5 %
0.7097121 0.7677752
```

$\hat{\beta}_1 = 0.739$ with a 95% interval of $(0.710, 0.768)$: each doubling of mass multiplies metabolic rate by $2^{0.739} = 1.67$, so metabolism rises less than proportionally with size. The interval contains Kleiber's 3/4 and excludes the surface-area value 2/3, so these data support Kleiber's law.

f. Test $H_0 : \beta_1 = 3/4$ and, separately, $H_0 : \beta_1 = 2/3$.

`summary()` tests $\beta_1 = 0$, which neither theory claims, so build the statistic yourself:

$$t = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{\text{se}[\hat{\beta}_1]}$$

```
b  <- coef(fit.k)['log(mass)']
se <- summary(fit.k)$coefficients['log(mass)', 'Std. Error']
c(vs.three.quarters = (b - 3/4) / se, vs.two.thirds = (b - 2/3) / se)
```

```
vs.three.quarters.log(mass)     vs.two.thirds.log(mass)
                 -0.7699518                   4.9301720
```

Against 3/4: $t = -0.77$, inside ± 1.96, consistent. Against 2/3: $t = 4.93$, outside, inconsistent. This matches (e) by construction: a value lies inside the 95% interval exactly when the test at $\alpha = 0.05$ fails to reject it.

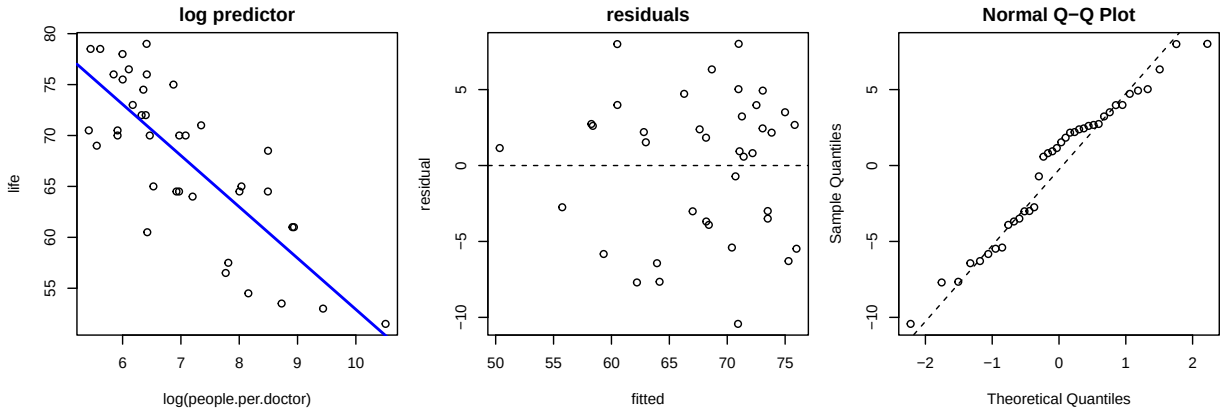
Life expectancy and access to doctors

4. The `doctors` data record life expectancy in years and the number of people per doctor for 38 countries in 1993.

   a. Identify and fit an appropriately specified model for the relationship between life expectancy and the number of people per doctor; write the model and report parameter estimates and standard errors.

The predictor spans 226 to 36,660 people per doctor and the scatterplot drops steeply then flattens, so a line in the raw predictor will not do. Logging the predictor straightens it:

```
fit.d <- lm(life ~ log(people.per.doctor), data = doctors)
par(mfrow = c(1, 3), mar = c(4, 4, 2, 1))
plot(life ~ log(people.per.doctor), data = doctors,
     main = 'log predictor')
abline(fit.d, col = 'blue', lwd = 2)
plot(fitted(fit.d), resid(fit.d), xlab = 'fitted', ylab = 'residual',
     main = 'residuals')
abline(h = 0, lty = 2)
qqnorm(resid(fit.d)); qqline(resid(fit.d), lty = 2)
```




```
summary(fit.d)$coefficients
```

```
                        Estimate Std. Error   t value     Pr(>|t|)
(Intercept)           103.28082  4.5207781 22.845806 5.122363e-23
log(people.per.doctor)  -5.03657  0.6316628 -7.973511 1.813580e-09
```

The fitted model is

$$\text{life}_i = \beta_0 + \beta_1 \log(\text{people per doctor})_i + \epsilon_i$$

with $\hat{\beta}_0 = 103.28$ (SE 4.52) and $\hat{\beta}_1 = -5.04$ (SE 0.63), and $\hat{\sigma} = 4.76$ on 36 degrees of freedom.

The diagnostics are clean: an even band of residuals, a QQ plot close to the line, largest Cook's distance 0.10. Only the predictor needed transforming; the spread of life expectancies is already even across the fitted range.

b. Report in context a 95% confidence interval for the parameter that measures the relationship.

```
confint(fit.d)['log(people.per.doctor)', ]
```

```
    2.5 %    97.5 %
-6.317642 -3.755499
```

```
log(2) * confint(fit.d)['log(people.per.doctor)', ]   # per doubling
```

```
    2.5 %    97.5 %
-4.379056 -2.603113
```

The interval for β_1 is $(-6.32, -3.76)$ years per unit of log(people per doctor). Multiplying by $\log 2$ makes it readable: each doubling of people per doctor is associated with a drop in mean life expectancy of between 2.6 and 4.4 years, point estimate 3.5.

c. Would you conclude from this fit that training more doctors raises life expectancy? Explain briefly.

No. These are observational data, and people per doctor travels with everything else that affects how long people live: income, nutrition, sanitation, education, public health infrastructure. Any of those could produce the association on its own. The fit supports association, not causation.