Handout 3

STAT3530 Applied Linear Models

This handout follows the lecture slides. This week we are covering inference and diagnostics in regression.

Review: inference for a population mean

Let $Y_1, \dots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$ and consider inference on the population mean μ.

1. The usual estimators are $\hat{\mu} =$ ______________ and $\hat{\sigma}^2 =$ ______________ .

2. The standard error for $\hat{\mu}$ is $SE(\hat{\mu}) =$ ______________ .

3. A two-sided test of $H_0 : \mu = \mu_0$ against $H_A : \mu \neq \mu_0$ uses the test statistic

$$T = \frac{\qquad - \qquad}{\qquad / \qquad}$$

which, under H_0, has a ______________ sampling distribution.

   a. The test rejects at level α when |_______| > _______ .

   b. The level α controls the t_______ ___ e_______ r_______ .

   c. A 95% confidence interval for μ is given by ________________________________.

   d. Supposing mean body temperature were estimated to be $\hat{\mu} = 98$ with standard error $SE(\hat{\mu}) = 1/1.96$ from a sample of $n = 200$ adults, compute and interpret a 95% confidence interval.

4. Write the model as a linear model.

Part 1: Inference for the coefficients

Now consider $Y = X\beta + \epsilon$ with $\mathbb{E}[\epsilon] =$ ______ and $\text{var}[\epsilon] =$ ______ .

From before we have $\hat{\beta} = (X^TX)^{-1}X^TY$ with $\text{Var}[\hat{\beta}] = \sigma^2(X^TX)^{-1}$, and $\hat{\sigma}^2 = \|e\|^2/(n-2)$.

5. Adding the normality assumption $\epsilon \sim N(0, \sigma^2 I)$ gives

$$\epsilon \sim N(0, \sigma^2 I) \implies Y \sim N(\underline{\hspace{3cm}}, \underline{\hspace{3cm}}) \implies \hat{\beta} \sim N(\underline{\hspace{3cm}}, \underline{\hspace{4cm}})$$

   a. So in particular, $\hat{\beta}_j \sim N(\underline{\hspace{3cm}}, \underline{\hspace{5cm}})$.

   b. Now the standard error of $\hat{\beta}_j$ is $SE(\hat{\beta}_j) =$ __________ .

   c. And we can construct a T statistic for β_j as

$$T_j = \underline{\hspace{4cm}} \sim t_{n-2}$$

6. To infer association between the predictor and mean response we test

$$H_0 : \underline{\hspace{3cm}} \quad \text{vs} \quad H_A : \underline{\hspace{3cm}}$$

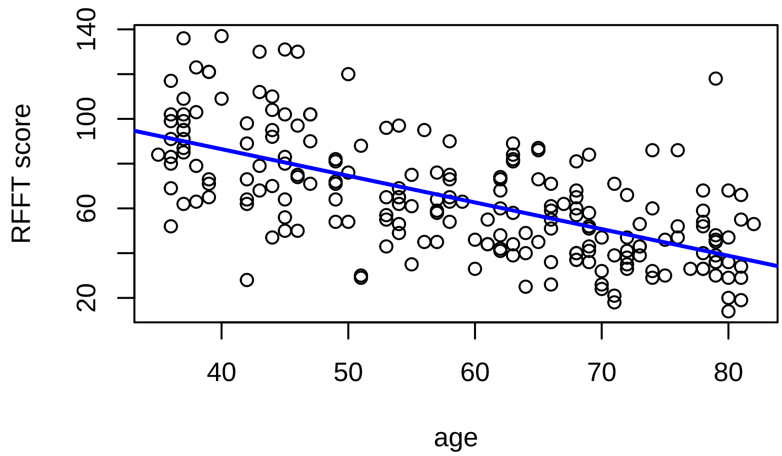
   using the test statistic

$$T = \underline{\hspace{4cm}}$$

   a. The data provide evidence of an association when…

   b. A __________% confidence interval for the relevant parameter is given by…

Recall the RFFT data: cognitive function score and age, for 208 adults.




```
# fit the model and check estimates
rfit <- lm(rfft ~ age, data = prevend)
summary(rfit)$coefficients
```

```
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 134.098052  6.0700593  22.09172 3.058596e-56
age          -1.190794  0.1007061 -11.82445 5.823319e-25
```

```
# confidence intervals
round(confint(rfit), 3)
```

```
              2.5 %  97.5 %
(Intercept) 122.131 146.065
age          -1.389  -0.992
```

7. Is cognitive function significantly associated with age? Highlight/circle the test statistic and p-value in the output above, and write a one-sentence conclusion.

8. Interpret the confidence interval for the slope.

Part 2: Prediction

Let $x_0 = [1 \quad x]^T$ be the predictor vector for a new observation and consider the prediction $\hat{Y}_0 = x_0^T\hat{\beta}$.

9. Show that $\text{Var}[\hat{Y}_0] = \sigma^2 h_{00}$ where $h_{00} = x_0^T(X^TX)^{-1}x_0$.

$$\text{Var}[x_0^T\hat{\beta}] = x_0^T \underline{\hspace{6cm}} x_0$$

$$= x_0^T\left[\underline{\hspace{7cm}}\right]x_0$$

$$= \sigma^2\, x_0^T(X^TX)^{-1}x_0$$

The same number $\hat{Y}_0$ can estimate two different things.

10. Fill in the table.

	the target	is it random?
mean response	$\mu_0 =$ ____________	__________
a new observation	$Y_0 =$ __________________	__________

11. For the mean, the error is $\hat{Y}_0 - \mu_0$. Explain why subtracting μ_0 does not change the variance, so that $\text{Var}[\hat{Y}_0 - \mu_0] = \text{Var}[\hat{Y}_0] = \sigma^2 h_{00}$.

12. For a new observation, write $Y_0 = x_0^T\beta + \epsilon_0$ and fill in the derivation.

$$\hat{Y}_0 - Y_0 = x_0^T\hat{\beta} - (x_0^T\beta + \epsilon_0) \;=\; x_0^T(\hat{\beta} - \beta) \;-\; \underline{\hspace{3cm}}$$

$$\text{Var}[\hat{Y}_0 - Y_0] = \underline{\hspace{5cm}} + \underline{\hspace{3.5cm}} \;=\; \sigma^2(1 + h_{00})$$

What has to be true for the second line to hold, and why is it true here?

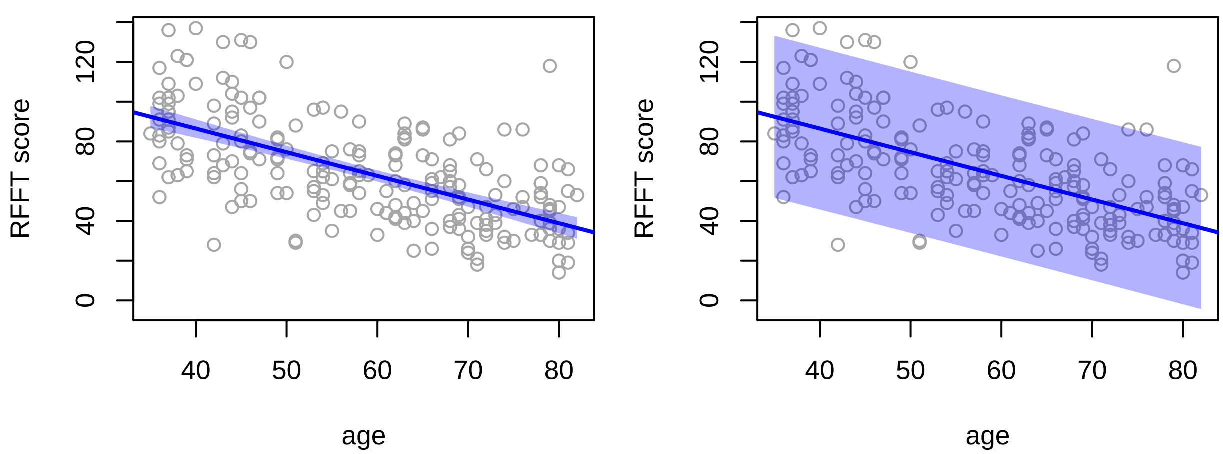
13. Write down both intervals and explain the difference in interpretation.

    confidence interval for μ_0: $\hat{Y}_0 \pm$ ________________________

    prediction interval for Y_0: $\hat{Y}_0 \pm$ ________________________

14. As $n \to \infty$ with the predictors spread over a fixed range, $h_{00} \to 0$. What happens to the width of each interval? Which one does *not* shrink to zero, and what does that mean in words?

15. The two bands below come from the same fit to the PREVEND data. Label which is which. At age 55 the two intervals are (65.71, 71.50) and (28.05, 109.16); interpret each in a sentence.



Part 3: Assumptions and diagnostics

16. List the four assumptions of $Y = X\beta + \epsilon$ in words and, to the right of each line, mathematically.

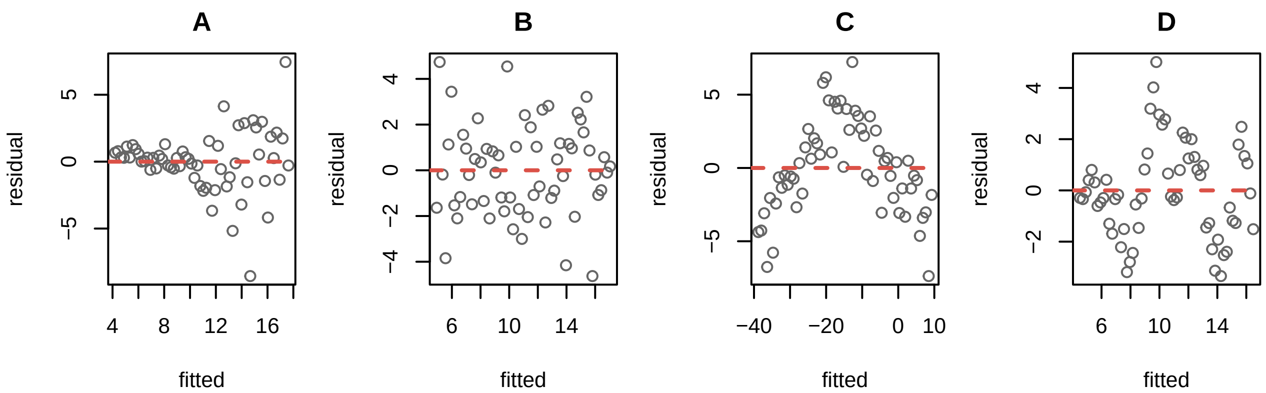
    1. ____
    2. ____
    3. ____
    4. ____

17. Sketch the two standard residual diagnostic plots and label the axes.

18. Complete the table.

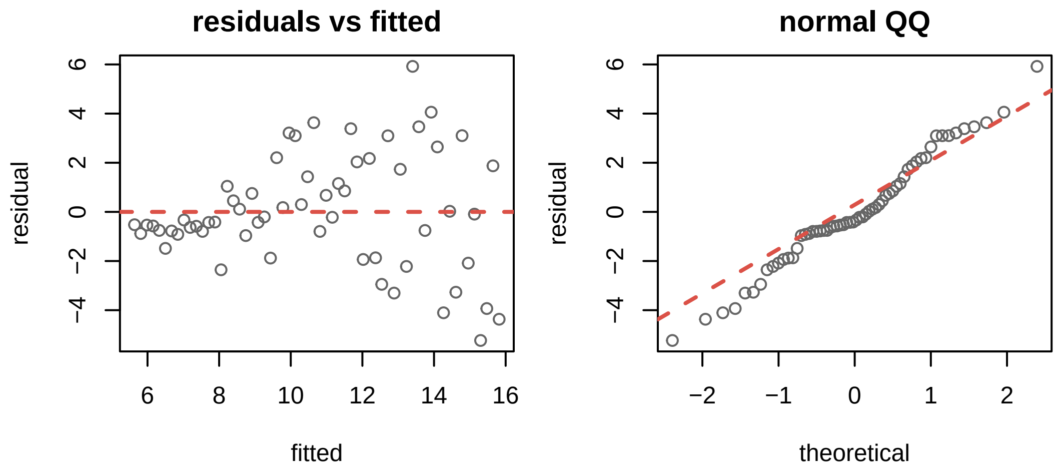
plot	assumption	what a violation looks like
residuals vs fitted	__________ and __________	
normal QQ	__________	

19. Each panel below is a residual plot from a different fit. Do the assumptions hold? If not, which one fails?



panel	do the assumptions hold?	if not, which one fails?
A		
B		
C		
D		

20. The two plots below come from the same fit.



a. Which assumption is violated, and which plot shows it?

b. Is $\hat{\beta}_1$ still unbiased here?

c. Is its reported standard error still right?

21. Below is the simulation from lecture: 2000 datasets, $n = 60$, true $\beta_1 = 3$.

errors	CI coverage	true SD	reported SE
well behaved	94.3	0.220	0.220
heavy tails	95.7	0.220	0.208
light tails	94.5	0.220	0.220
variance, pronounced	87.8	0.562	0.428
correlation, pronounced	47.0	0.584	0.190

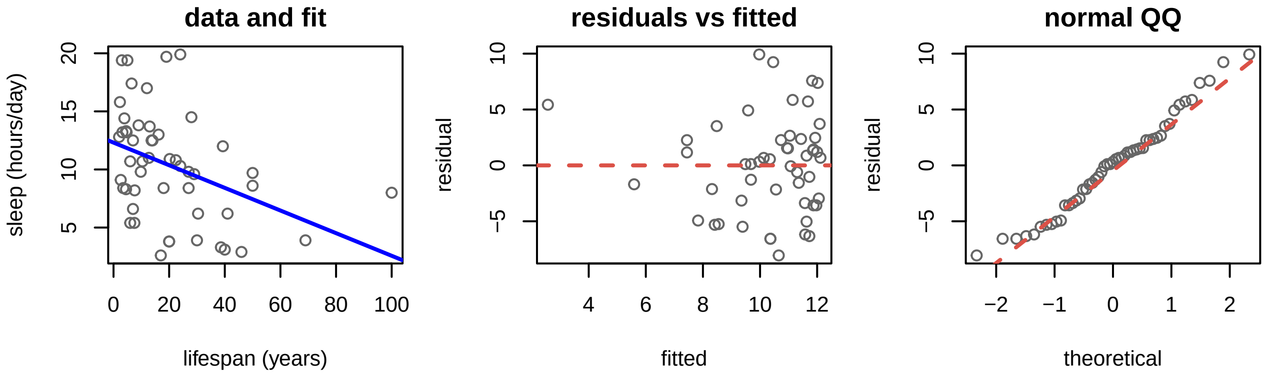
(a) Which settings make inference unreliable? Put stars next to the rows.

(b) In the last row, the reported standard error is three times too small. Explain the consequence for a 95% interval.

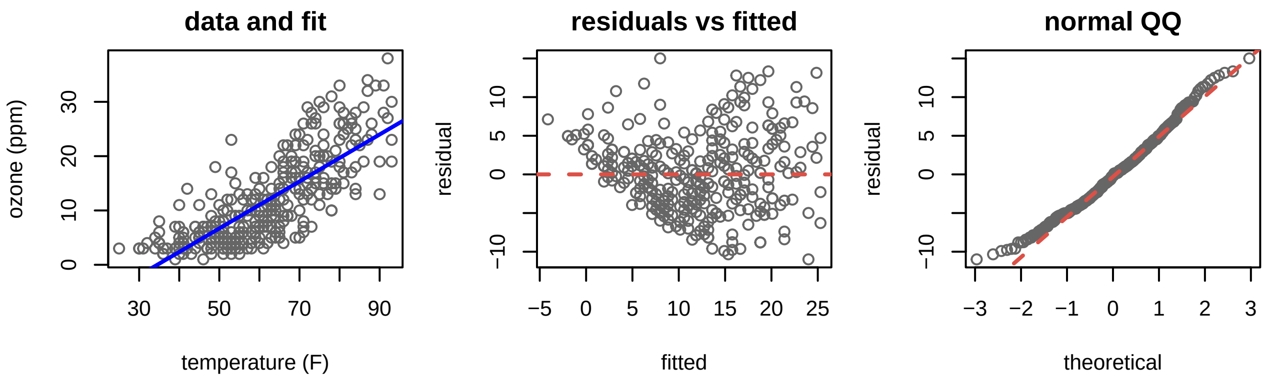
22. Rank the four model assumptions in approximate order of how much a violation matters.

Examples. For each fit below, decide whether the assumptions hold, and name any that fail.

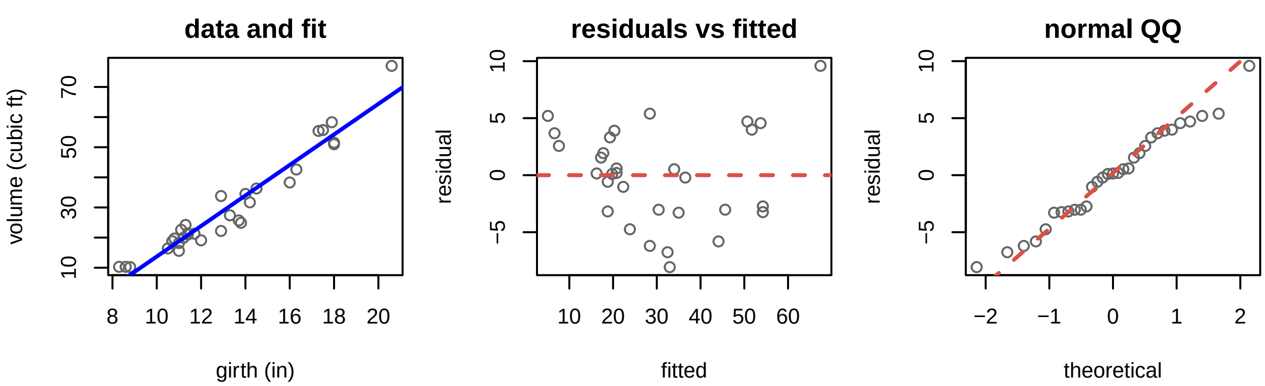
Mammal sleep. Hours of sleep per day against lifespan, for 51 species.



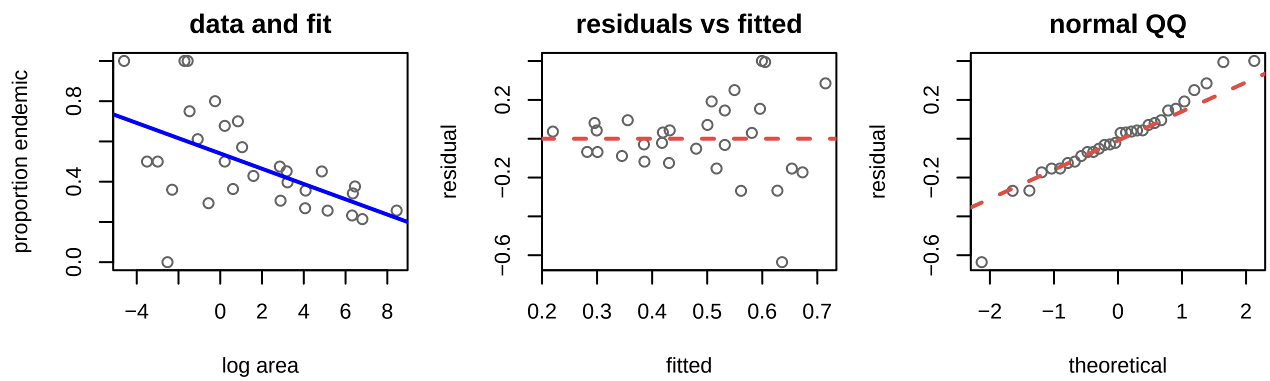
Ozone and temperature. Daily ozone against temperature in the Los Angeles basin, 330 days.



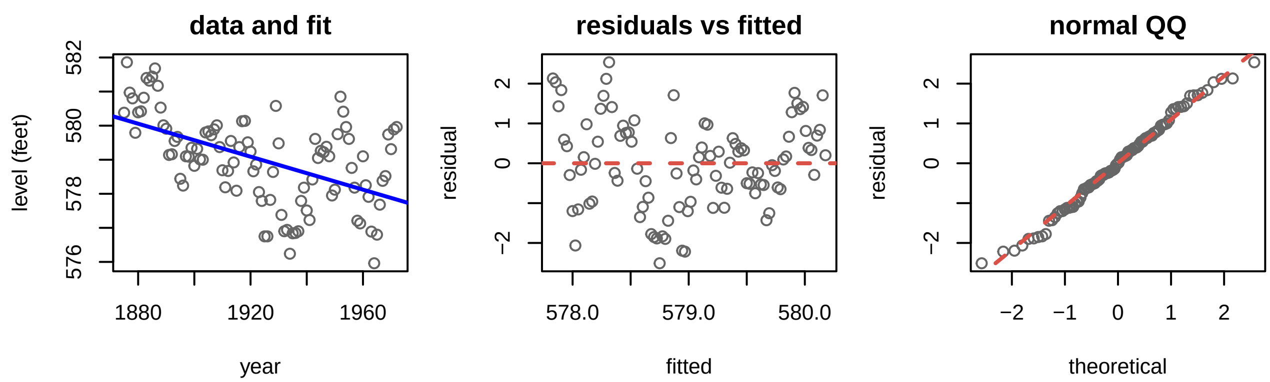
Black cherry trees. Timber volume against trunk girth for 31 felled black cherry trees.



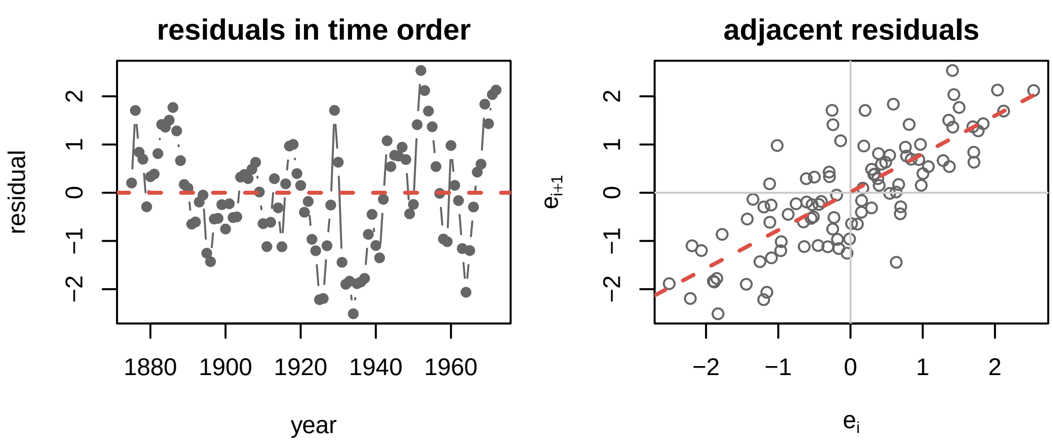
Endemic species. Proportion of each Galapagos island's species that are endemic, against log area.



Lake Huron. Annual level of Lake Huron, 1875–1972, regressed on year.



Adjacent years might be correlated, and that is invisible in the plots above. Put the residuals in time order and check.



Part 4: Transformations

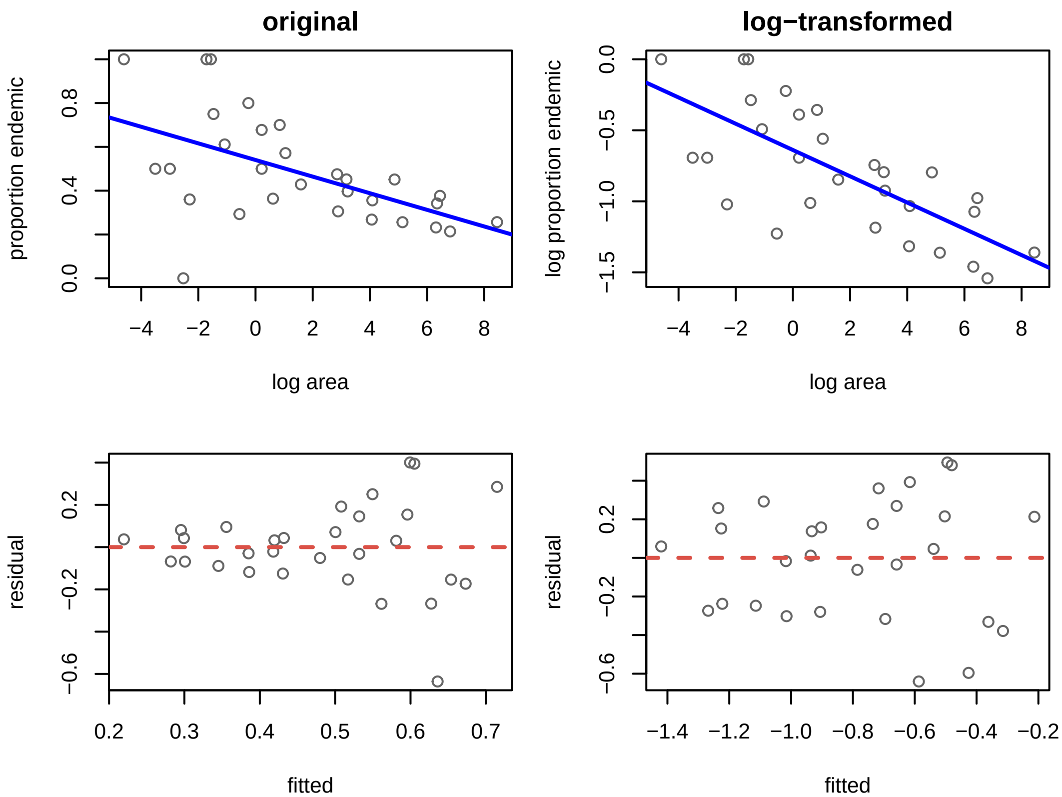
For the Galapagos data we modelled the proportion of endemic species against log area, and found the spread of the residuals grew with the fitted value. Taking logs of the response fixes most of it.

Original model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

After log-transforming the response:

$$\log(Y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$$



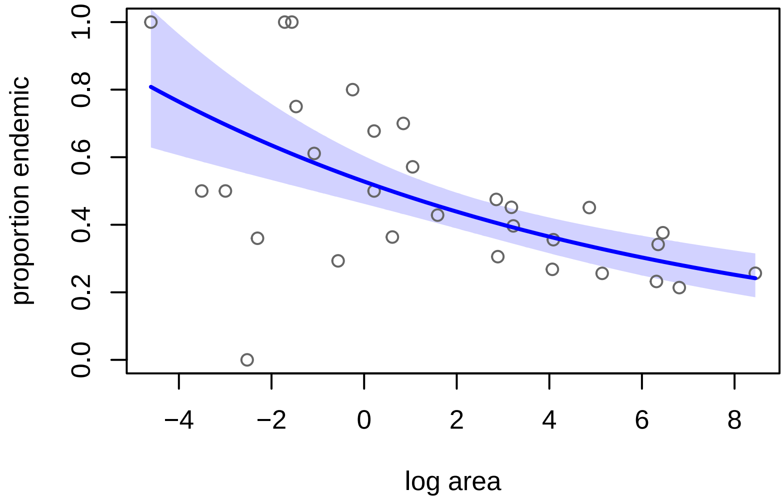
23. Starting from $\log(Y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$, exponentiate both sides to write Y_i as a product. What has happened to the errors on the original scale?

$$Y_i \;=\; \underline{\hspace{6cm}}$$

24. If $\log(Y) \sim N(\mu, \sigma^2)$ then e^{μ} is the **median** of Y, not the mean. Using that, complete each statement.

- median$(Y_i) =$ ______________________
- e^{β_0} is the median when _________________
- increasing x_i by one changes the median by a factor of __________

Back-transforming the fit to the original scale:




```r
# fit on transformed scale
efit <- lm(log(prop) ~ la, data = endemics, subset = keep)
summary(efit)$coefficients
```

```
               Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -0.63875781 0.06513603 -9.806521 2.155528e-10
la          -0.09246658 0.01708572 -5.411920 1.009005e-05
```

25. The fitted model gives $\hat{\beta}_1 = -0.0925$, so $e^{\hat{\beta}_1} = 0.912$, with interval $(0.88,\ 0.944)$. Interpret the point estimate and interval in context.

26. Here the predictor is *also* logged, so a one-unit change in x is tricky to carry through the model. A more natural change is a doubling of x.

    a. Rewrite the model in terms of Y_i and x_i (not $\log(x_i)$).

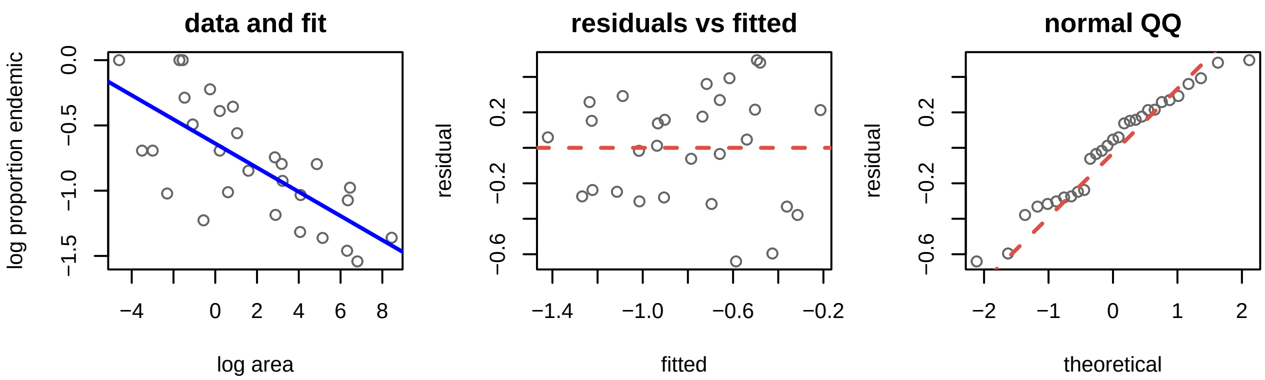
    b. Show that doubling area multiplies the median by 2^{β_1}.

$$\beta_1 \log(2x) = \beta_1[\underline{\hspace{2cm}} + \underline{\hspace{2cm}}] = \underline{\hspace{2cm}} + \underline{\hspace{2cm}}$$

$$\exp\{\beta_1 \log(2x)\} = \underline{\hspace{5cm}}$$

b. Compute a point estimate for the multiplicative change in median share of endemic species per doubling of island area.

The assumptions live on the transformed scale, so that is where the diagnostics belong.



27. Comment on the residual diagnostics. Do the assumptions seem plausible?