Handout 4

STAT3530 Applied Linear Models

This handout follows the lecture slides. This week we are covering transformations and unusual observations.

Review: troubleshooting

1. Once you identify an issue with a model, there are three options. Fill them in, along with an issue each one is suited to.

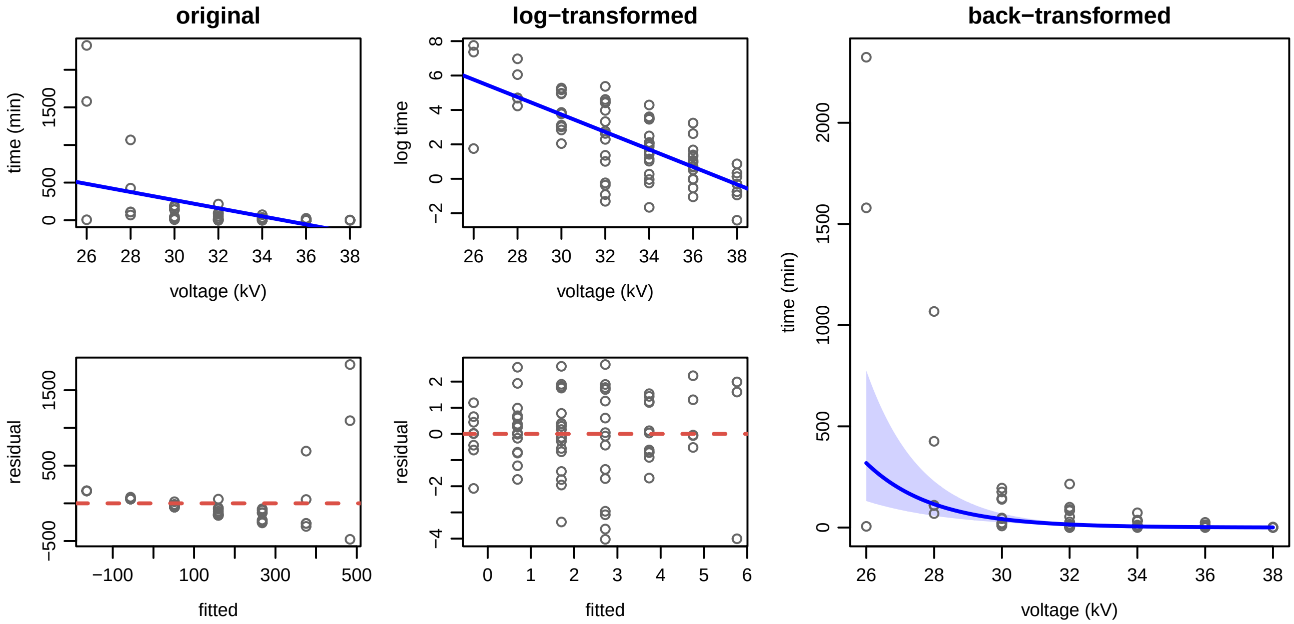
option	suited to
1. ____________________	__________________________
2. ____________________	__________________________
3. ____________________	__________________________

2. After transforming, the model assumptions apply on the _______________ scale, so that is where the diagnostics should be checked.

Part 1: Log transformations

The response

Time until an electrical insulating fluid breaks down, for 76 samples tested at seven voltages.




```r
fit <- lm(log(time) ~ voltage, data = fluid)
summary(fit)$coefficients
```

```
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 18.9554587 1.91001876  9.924226 3.052326e-15
voltage     -0.5073649 0.05739595 -8.839734 3.340072e-13
```

3. Using the plots and output above:

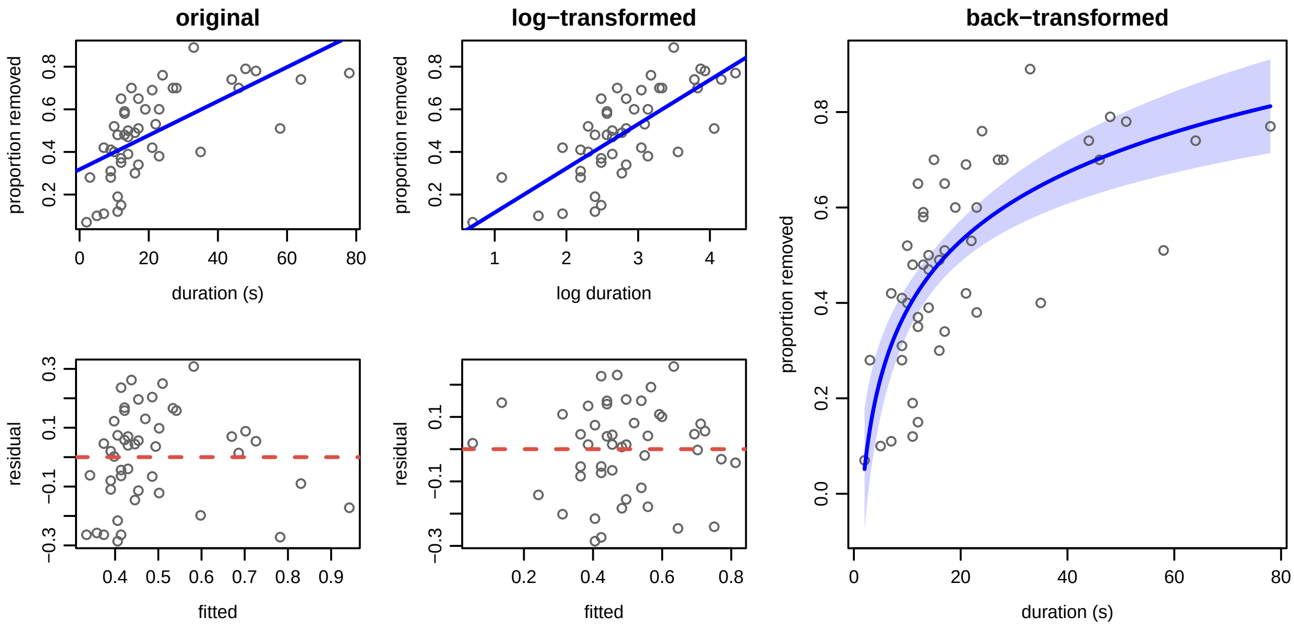
   a. The issue with the original fit is ______________________________.

   b. The fitted model on the original scale is $\hat{Y}_i = e^{\_\_\_\_\_\_} e^{\_\_\_\_\_\_ x_i}$, and $\hat{Y}_i$ estimates the __________ breakdown time.

   c. Compute the factor by which the median breakdown time changes for each additional kV, and interpret it in context.

The predictor

Proportion of a lily's pollen removed by a bee, against how long the visit lasted, for 47 visits.




```r
fit <- lm(removed ~ log(duration), data = pollen)
summary(fit)$coefficients
```

```
                 Estimate Std. Error   t value     Pr(>|t|)
(Intercept)   -0.09221658 0.08149970 -1.131496 2.638404e-01
log(duration)  0.20760092 0.02822562  7.355052 3.040327e-09
```

4. Using the plots and output above:

   a. The issue with the original fit is ____.

   b. The fitted model is $\hat{Y}_i =$ ________ $+$ ________ $\log(x_i)$, and $\hat{Y}_i$ estimates the ___________ proportion removed.

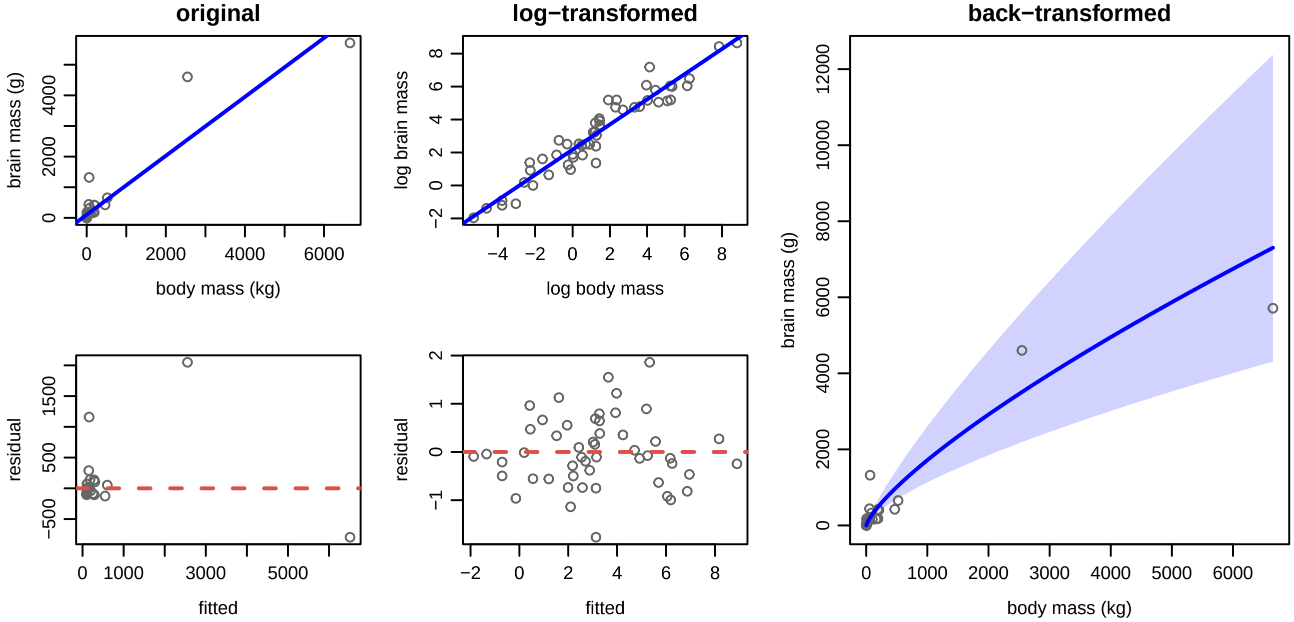
   c. Show that doubling the length of a visit adds $\beta_1 \log(2)$ to the mean.

   $$\beta_0 + \beta_1 \log(2x) = \beta_0 + \beta_1 [\_\_\_\_\_\_\_\_ + \_\_\_\_\_\_\_\_] = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

   d. Compute the change in mean proportion removed for each doubling of visit length, and interpret it in context.

Both

Brain mass against body mass for 51 species of mammal.




```r
fit <- lm(log(brain) ~ log(body), data = mammals)
summary(fit)$coefficients
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 2.1737368 0.11018768 19.72758 5.870654e-25
log(body)   0.7636132 0.03234614 23.60755 2.022309e-28
```

5. Using the plots and output above:

   a. Exponentiate $\log(Y_i) = \beta_0 + \beta_1 \log(x_i) + \epsilon_i$ to write the model in terms of Y_i and x_i.

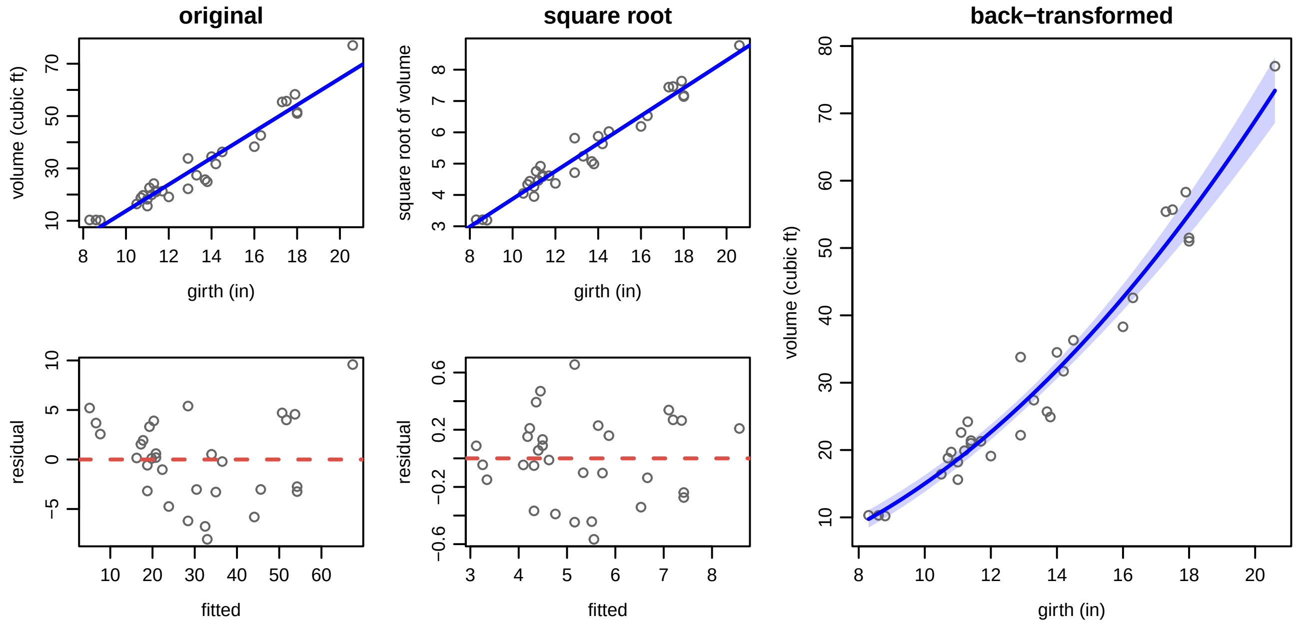
   $$Y_i = \underline{\hspace{6cm}}$$

   b. Compute the factor by which median brain mass changes for each doubling of body mass, and interpret it in context.

Part 2: Power transformations

Square root

Timber volume against trunk girth for 31 black cherry trees.




```r
fit <- lm(sqrt(Volume) ~ Girth, data = trees)
summary(fit)$coefficients
```

```
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -0.5518275 0.23718993 -2.326522 2.718265e-02
Girth        0.4426244 0.01743619 25.385381 2.347333e-21
```

7. Using the plots and output above:

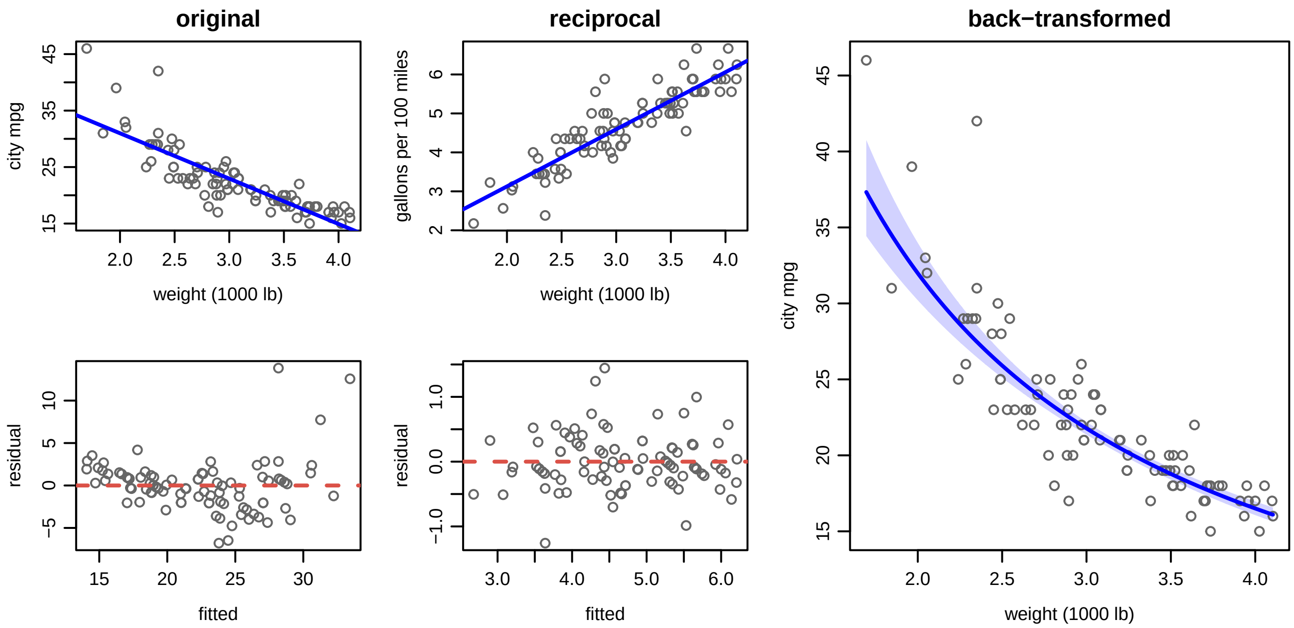
   a. Volume grows roughly with girth squared. Why does that suggest taking the square root of volume?

   b. The fitted model on the original scale is $\hat{Y}_i = (\underline{\hspace{3em}} + \underline{\hspace{3em}}\, x_i)^2$.

   c. Compute the estimated median volume at girths of 10 and 11 inches, and again at 18 and 19 inches. Why does an extra inch matter more for a big tree?

Reciprocal

City fuel economy against weight for 93 car models on sale in the US in 1993.




```r
fit <- lm(100/mpg ~ weight, data = cars93)
summary(fit)$coefficients
```

```
             Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 0.1936667 0.23708840  0.8168544 4.161450e-01
weight      1.4662292 0.07578551 19.3470924 5.265735e-34
```

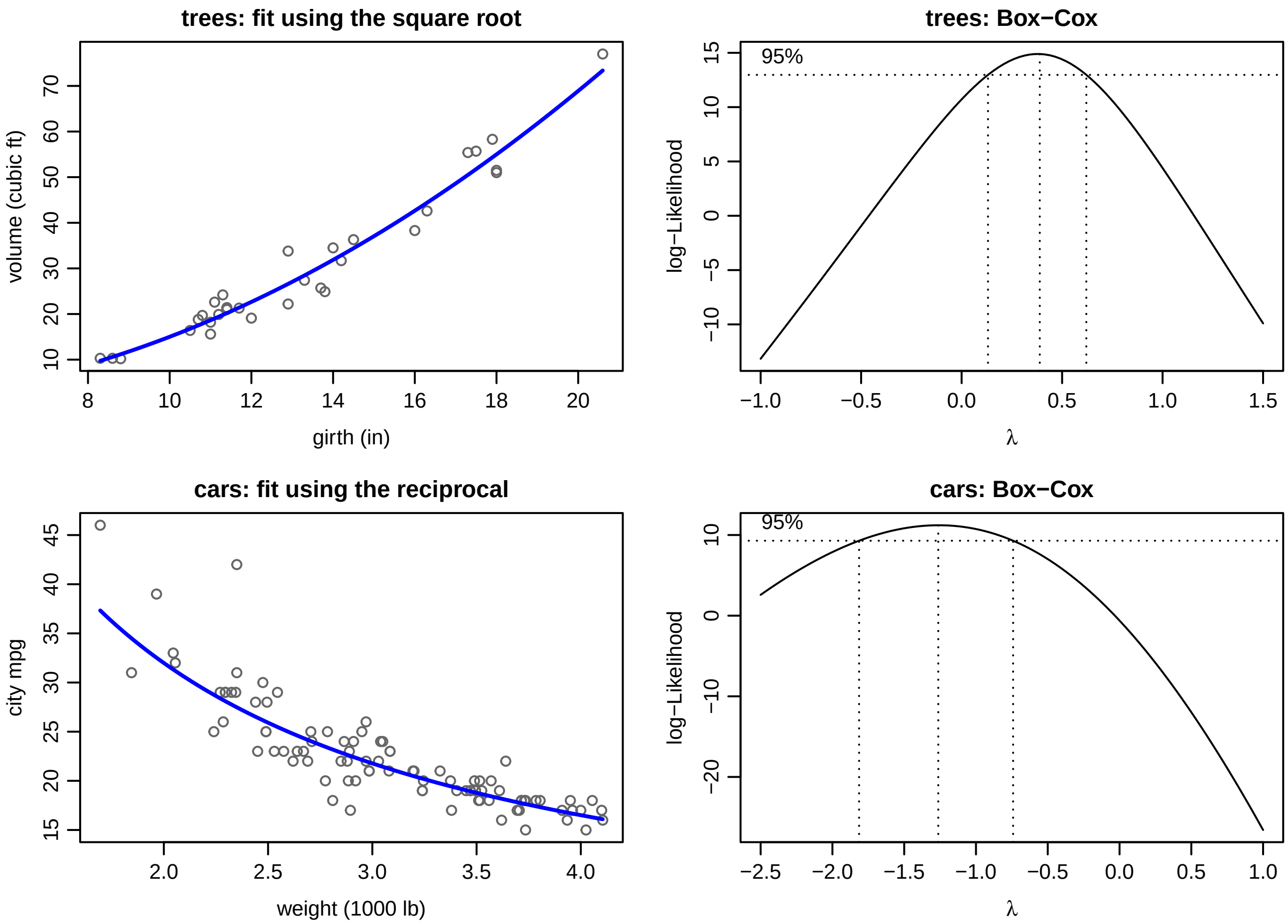
8. Using the plots and output above:

    a. Fuel economy Y is measured in miles per gallon. What does $100/Y$ measure?

    b. Interpret the estimated slope in context.

    c. Compute the estimated median fuel economy for a car weighing 3000 lb.

Box-Cox



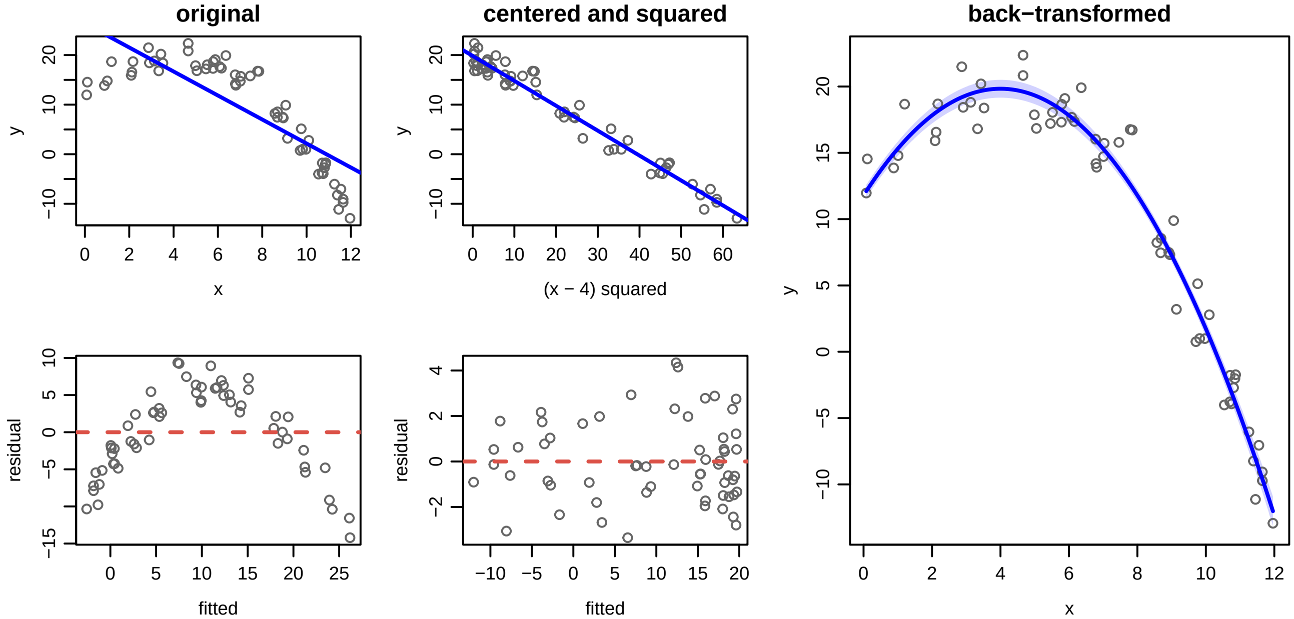
9. The Box-Cox method tries many ____________ of the response and scores how well the model fits for each one. The ____________ score marks the best power, and powers that score almost as well are also reasonable.

10. Read the plots above.

	best power (approx.)	reasonable range	simple power to use
trees	__________	______________	__________
cars	__________	______________	__________

Centering and squaring the predictor

Some relationships rise and then fall, or fall and then rise. Below are 60 simulated observations whose trend turns around near $x = 4$.




```r
fit <- lm(y ~ I((x - 4)^2), data = hump)
summary(fit)$coefficients
```

```
               Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  19.8322416 0.33641317  58.95204 1.835915e-53
I((x - 4)^2) -0.5027307 0.01175219 -42.77763 1.447942e-45
```

11. Using the plots and output above:

    a. Why can't a log or power transformation straighten this relationship?

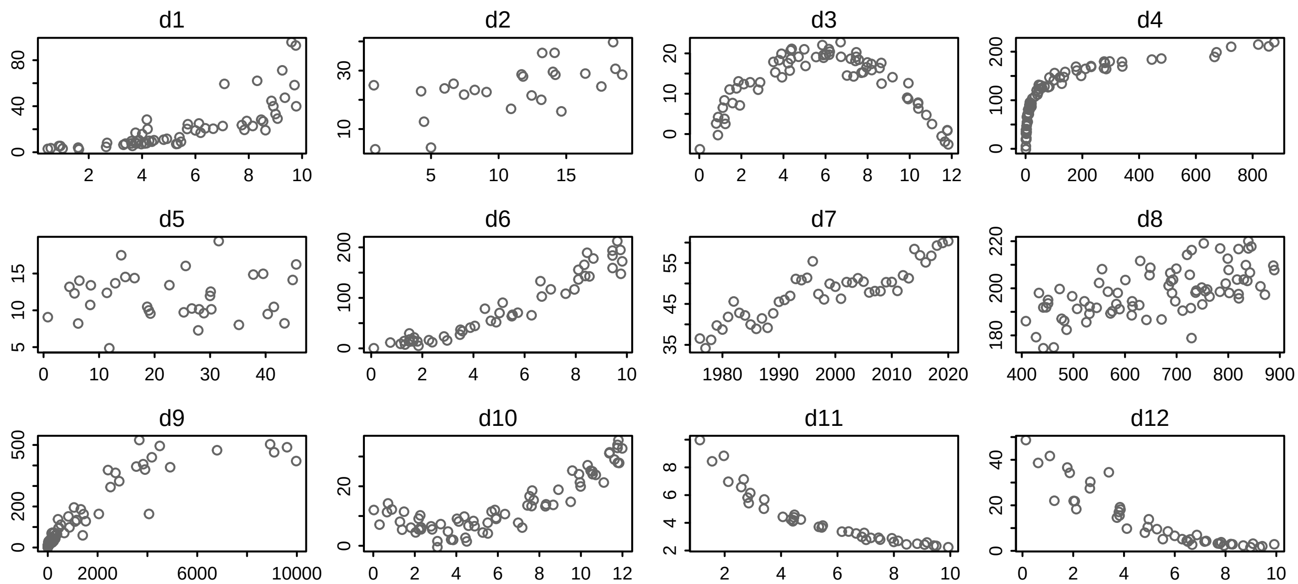
    b. After centering the predictor at c and squaring, the model is $Y_i = \beta_0 + \beta_1(\underline{\hspace{3cm}})^2 + \epsilon_i$.

    c. The fitted model is $\hat{Y}_i = \underline{\hspace{2cm}} + \underline{\hspace{2cm}}(x_i - 4)^2$. Where does the mean response peak, and what is the estimated peak?

    d. If the turning point were hard to see, how could you choose c? *(Hint: the response is the same for every choice of c.)*

Activity: diagnose and fix

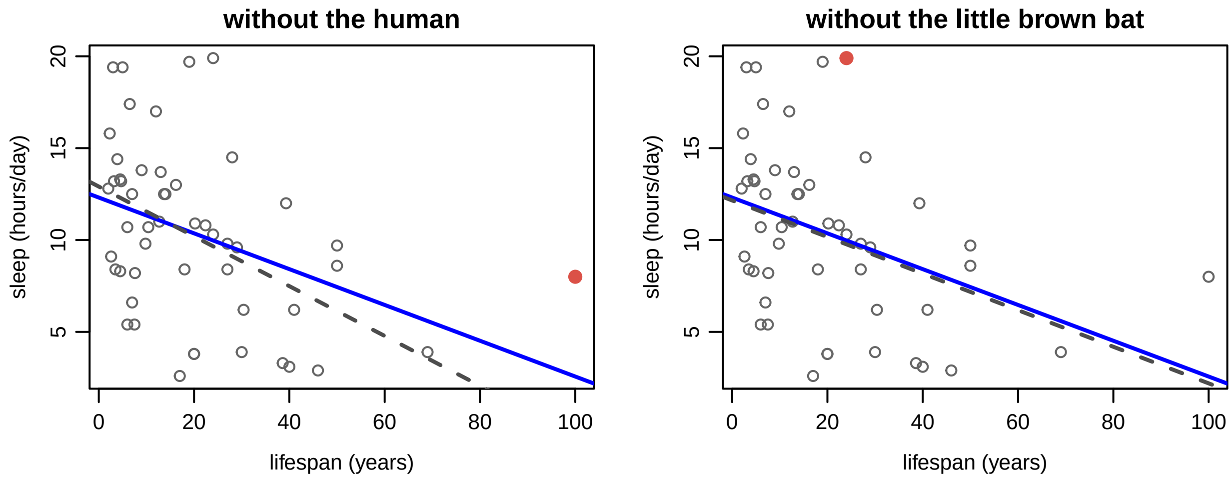
Twelve simulated datasets, each with one predictor. For each one, diagnose the problem, try a fix, and check that the fix worked. The script is `scripts/06-diagnose-and-fix.R` on the course site.



data	issue, or "none"	fix	clean afterwards?
d1			
d2			
d3			
d4			
d5			
d6			
d7			
d8			
d9			
d10			
d11			
d12			

Part 4: Unusual observations

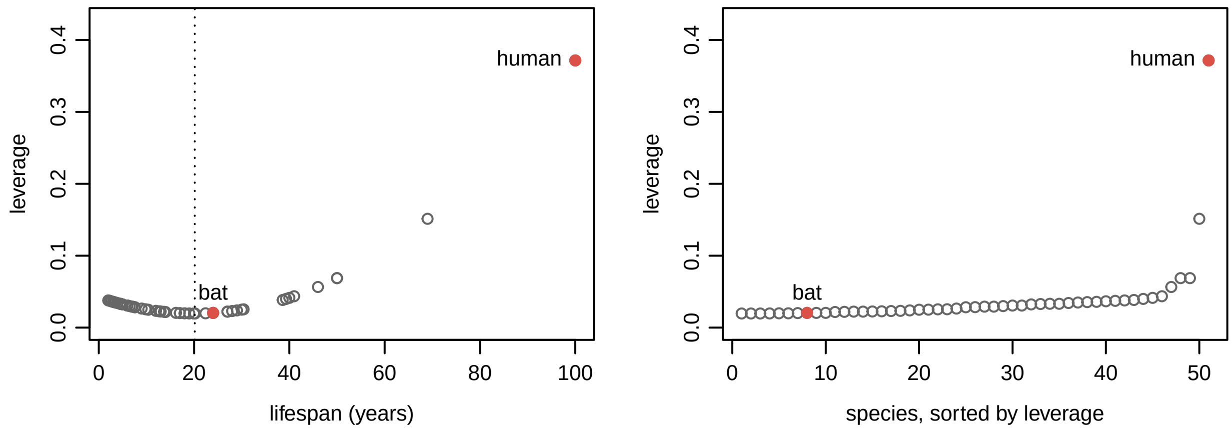
Hours of sleep per day against lifespan for 51 species. Solid line: all the data. Dashed line: without the red point.



12. Fill in the table.

term	what makes a point unusual	measured by
leverage	______	______
outlier	______	______
influential point	______	______

Leverage

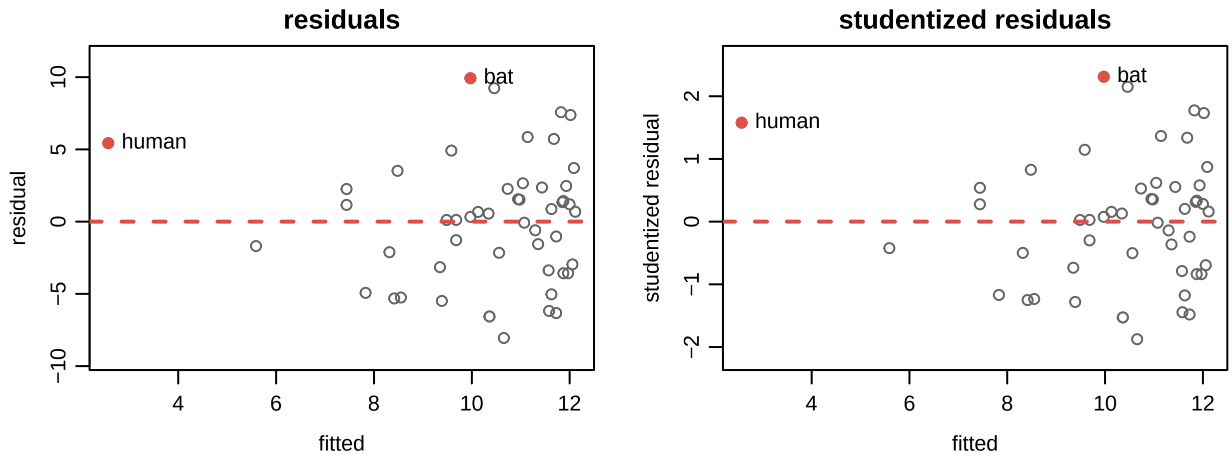


13. For simple linear regression, leverage is a squared distance from $\bar{x}$, measured in standard deviations:

$$h_{ii} = \frac{1}{n} + \frac{1}{n-1}\left(\qquad\qquad\qquad \right)^2$$

Humans live 4.2 standard deviations longer than the average species here, and $n = 51$. Compute their leverage.

Studentized residuals

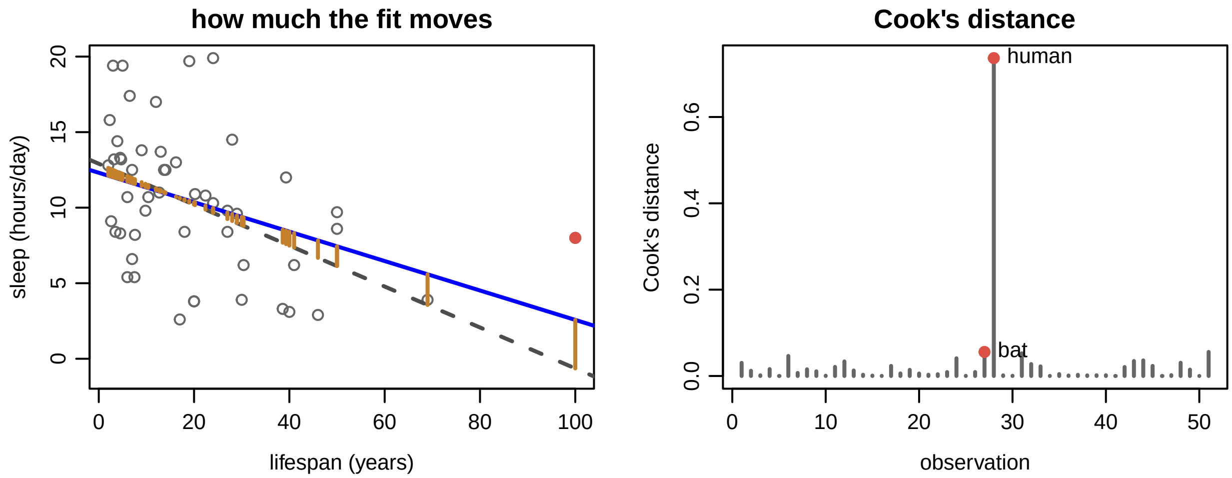


14. Since $e = (I - H)\epsilon$, the variance of a residual depends on leverage:

$$\mathrm{var}[e_i] = \sigma^2 \underline{\hspace{3cm}} \qquad \text{so we define} \qquad r_i = \frac{e_i}{\hspace{4cm}}$$

The two plots use the same vertical scale. Why does the human move up in the second plot while the bat barely moves?

Cook's distance



15. Cook's distance measures how much *all* the fitted values move when observation i is removed:

$$D_i = \frac{\sum_j \left(\qquad\qquad\qquad\qquad \right)^2}{p\,\hat{\sigma}^2} = \frac{r_i^2}{p} \cdot \left(\qquad\qquad\qquad \right)$$

a. The orange segments in the first plot show $\hat{Y}_j - \hat{Y}_{j(i)}$ for the human. Why are they longest on the right?

b. With $p = 2$, compute D_i for the human ($r = 1.58$, $h = 0.372$) and for the bat ($r = 2.31$, $h = 0.020$). Which is influential, and why?

With and without humans

```r
# all 51 species
round(summary(sleep.fit)$coefficients, 4)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.3117     0.8896 13.8399    0.000
lifespan     -0.0974     0.0322 -3.0212    0.004
```

```r
# without humans
no.human <- lm(sleep ~ lifespan, data = mammals[-human, ])
round(summary(no.human)$coefficients, 4)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.9095     0.9517 13.5641   0.0000
lifespan     -0.1355     0.0396 -3.4179   0.0013
```

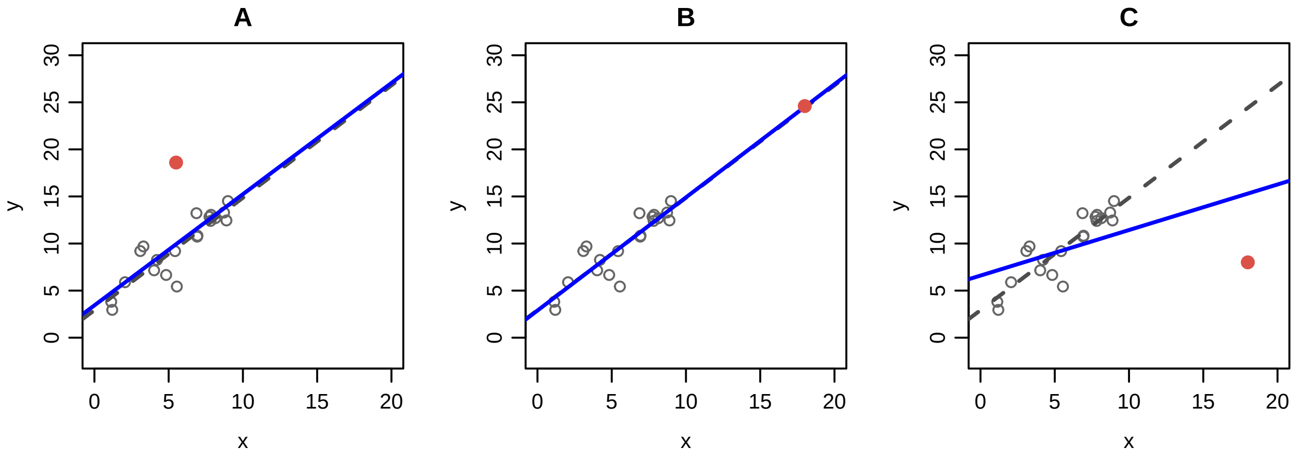
16. Using the output above:

    a. Does removing humans change whether lifespan is significantly associated with sleep?

    b. By how many standard errors of the all-species fit does the slope move?

    c. For which species would predictions change the most, and why?

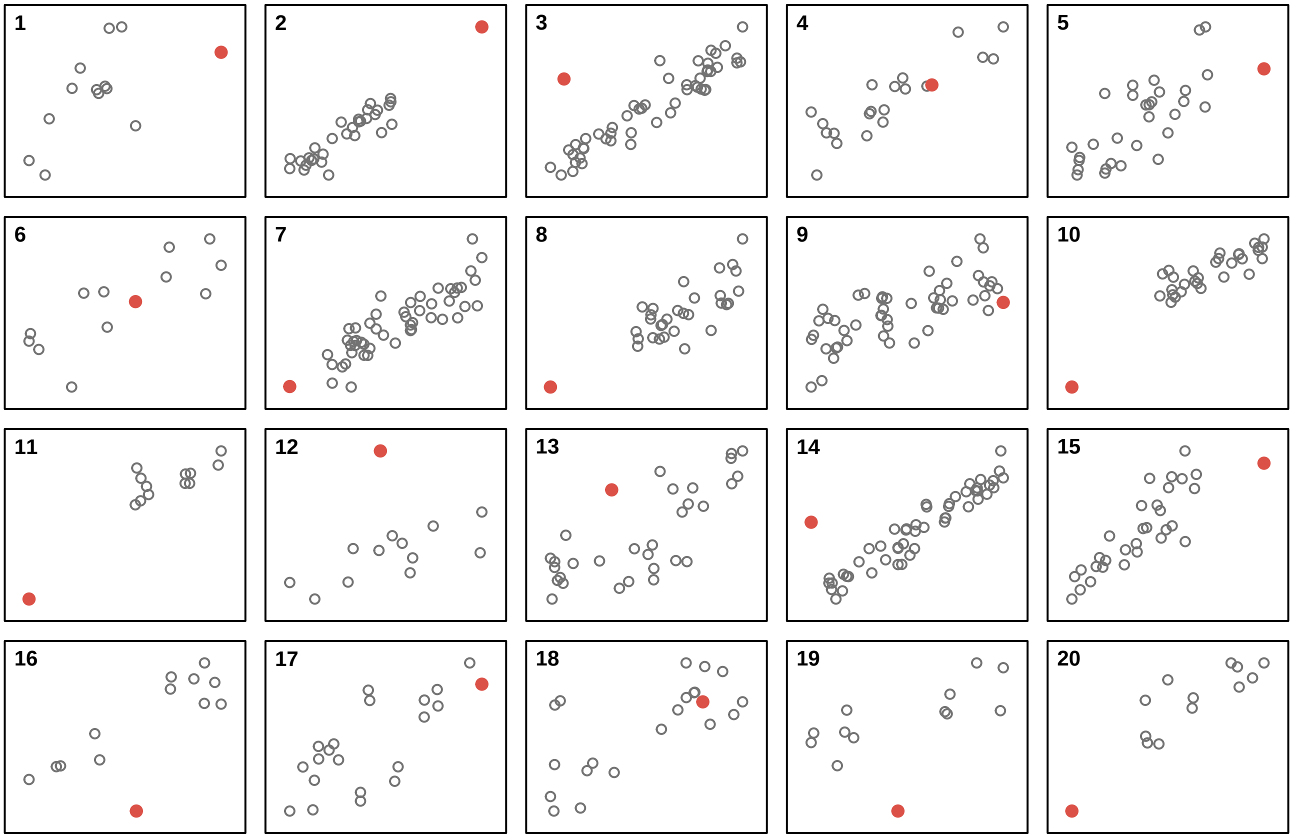
Three kinds of unusual



17. For each panel, is the red point high leverage, an outlier, and influential?

panel	high leverage?	outlier?	influential?
A			
B			
C			

Now you guess



18. In each panel, mark the red point **L** if it has high leverage and **O** if it is an outlier. Circle the panels where you think it is influential, then check your guesses against the slides.