Midterm Exam: Practice Set

STAT3530 Applied Linear Models

This set mirrors the structure of the exam: five questions, each with a handful of short parts. Work it under exam conditions if you can — open note, a calculator, no software — and give yourself about an hour.

1. A lab takes $n = 5$ independent measurements of the same quantity, so that

$$Y_i = \mu + \epsilon_i, \qquad \mathbb{E}[\epsilon] = 0, \qquad \text{Var}[\epsilon] = \sigma^2 I$$

A colleague claims the second measurement of the day is the only one worth keeping, and proposes estimating μ by

$$\tilde{\mu} = Y_2$$

a. Write $\tilde{\mu} = a^T Y$ by giving the vector a.

b. Is $\tilde{\mu}$ unbiased for μ? Recall $\mathbb{E}[a^T Y] = a^T \mathbb{E}[Y]$.

c. The least squares estimator of μ is $\hat{\mu} = \bar{Y}$. Is this a better or worse estimator than $\tilde{\mu}$? Explain.

2. An instrument returns a reading Y on a specimen of known concentration x. The instrument reads zero on a blank specimen, so the model carries no intercept:

$$Y_i = \beta x_i + \epsilon_i, \qquad \mathbb{E}[\epsilon] = 0, \qquad \mathrm{Var}[\epsilon] = \sigma^2 I$$

In matrix form $Y = X\beta + \epsilon$, where X is the single column x.

a. Find $X^T X$ and $X^T Y$, writing each as a sum.

b. Find $(X^T X)^{-1}$, and use it to show that

$$\hat{\beta} = \frac{\sum_i x_i Y_i}{\sum_i x_i^2}$$

c. Using $\mathrm{Var}[\widehat{\beta}] = \sigma^2 (X^T X)^{-1}$, write down $\mathrm{Var}[\widehat{\beta}]$. What happens to it as the concentrations are spread further from zero?

The lab runs $n = 12$ specimens and obtains

$$\sum_i x_i^2 = 650, \qquad \sum_i x_i Y_i = 1560, \qquad \widehat{\sigma} = 0.80$$

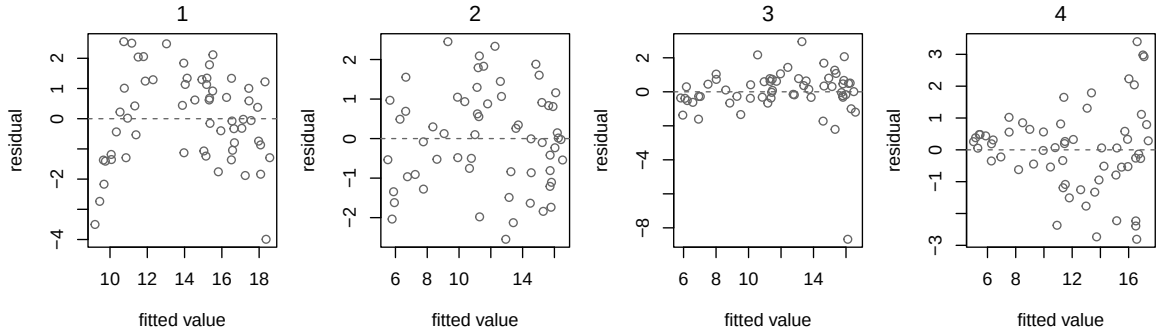
with $t_{0.975} = 2.201$ on 11 degrees of freedom.

d. Compute $\widehat{\beta}$ and a 95% confidence interval for β.

e. What would you check in the data before trusting a model with no intercept?

3. Spot checks.

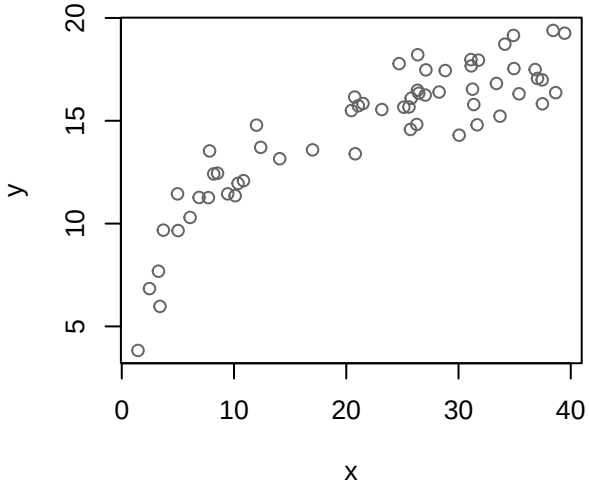
   a. Each residual plot below comes from a different fitted simple linear model. Name the problem each one shows, or write "none".



   1: _____ 2: _____

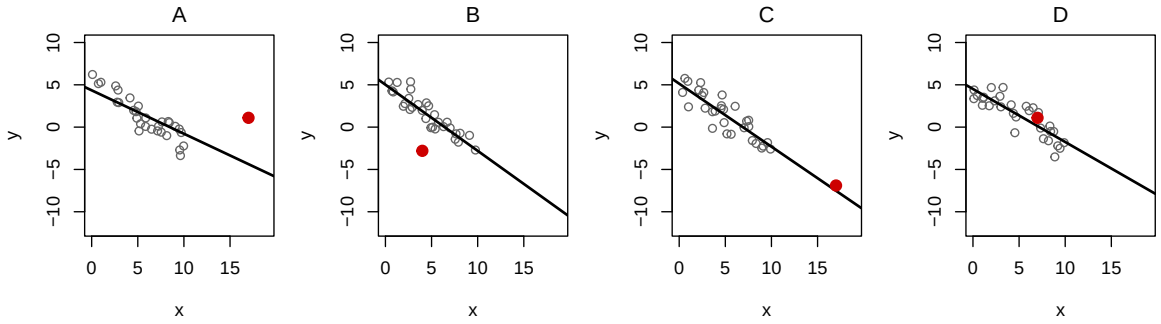
   3: _____ 4: _____

   b. For each panel that shows a problem, propose a fix. The scatterplot for panel 1 is below; note that the spread of the points about the curve does not change from left to right.



c. For each panel, identify tasks that the model *could* reliably support in spite of any issues with model assumptions, or state "none".

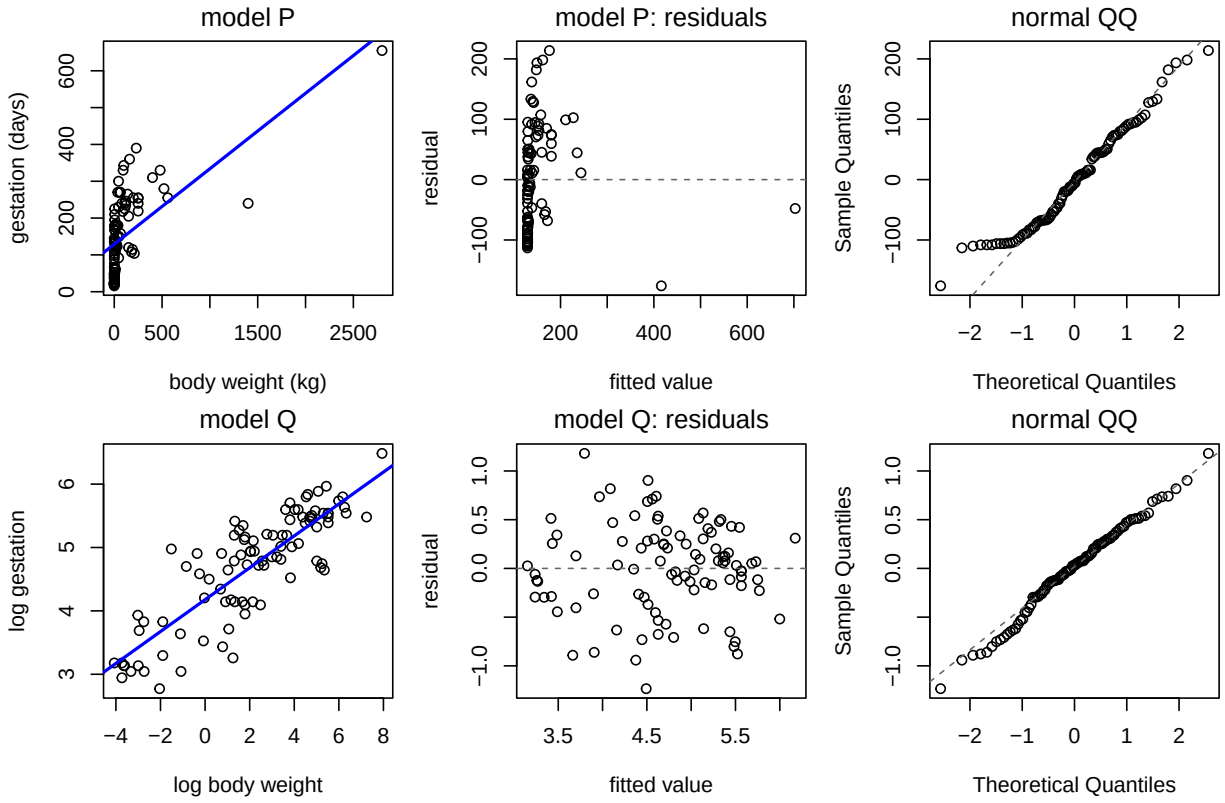
d. Each plot below shows a fitted simple linear model with one observation marked. Label each panel according to whether the marked point is **high leverage (L)**, **an outlier (O)**, **influential (I)**, or **none of these (N)**. A point can be any combination of L, O, and I.



A: _____ B: _____

C: _____ D: _____

4. Larger mammals carry their young longer, and a zoologist wants a usable description of that relationship: how much longer is gestation for a heavier species? The data are the body weight, in kg, and the gestation length, in days, of 96 mammal species. Two fits are shown.




```
summary(fitP)$coefficients
```

```
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 129.0608671 9.11926890 14.152545 5.011004e-25
Body          0.2049295 0.02642252  7.755866 1.035060e-11
```

```
summary(fitQ)$coefficients
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 4.1772145 0.05891226 70.90569 2.056069e-83
log(Body)   0.2515955 0.01638662 15.35371 2.335952e-27
```

Model P has $\hat{\sigma} = 84.8$; model Q has $\hat{\sigma} = 0.465$. Both are on 94 degrees of freedom, and $t_{0.975,94} = 1.986$.

a. Which model would you use to answer the zoologist's question, and why?

b. Using the model you chose, report a 95% confidence interval for the slope and interpret it in context. What happens to gestation when a species is twice as heavy?

c. A species weighs 50 kg. Estimate its gestation length, and report the interval below that is appropriate for a single new species.

```
predict(fitP, data.frame(Body = 50), interval = 'prediction')
```

```
       fit       lwr      upr
1 139.3073 -30.03704 308.6517
```

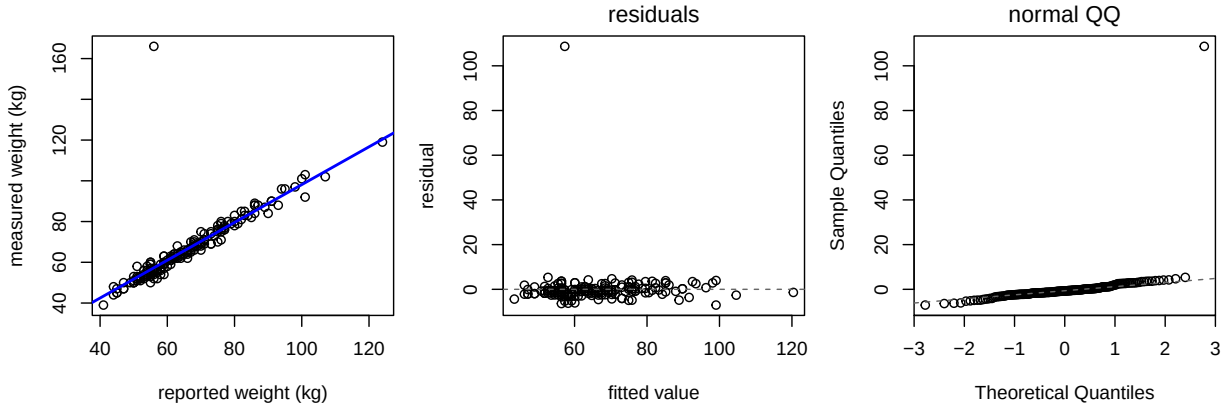
```
exp(predict(fitQ, data.frame(Body = 50), interval = 'confidence'))
```

```
       fit      lwr      upr
1 174.4192 156.1536 194.8215
```

```
exp(predict(fitQ, data.frame(Body = 50), interval = 'prediction'))
```

```
       fit      lwr      upr
1 174.4192 68.83375 441.9645
```

5. Surveys often ask people to report their own weight rather than weighing them. If self-reports were accurate on average, then measured weight would track reported weight one for one — a one-kilogram difference in what people report would correspond to a one-kilogram difference in what the scale says. The data below are the measured and self-reported weights, in kg, of 183 adults.




```
summary(fitW)
```

```
Call:
lm(formula = weight ~ repwt, data = wt)

Residuals:
    Min      1Q  Median      3Q     Max
 -7.048  -1.868  -0.728   0.601 108.705

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.3363     3.0369   1.757   0.0806 .
repwt         0.9278     0.0453  20.484   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.419 on 181 degrees of freedom
Multiple R-squared:  0.6986,    Adjusted R-squared:  0.697
F-statistic: 419.6 on 1 and 181 DF,  p-value: < 2.2e-16
```

a. Is the model adequately specified for inference? Comment on the diagnostic plots.

b. Are self-reports accurate on average? State the hypotheses that answer this question, and build the test statistic yourself from the estimate and its standard error; `summary()` does not test it for you. Compare with the critical value 1.97 and state your conclusion.

c. One subject is flagged in the diagnostics. Refitting without that subject gives the output below. Redo the test in (b), and say whether your conclusion changes.

```
summary(lm(weight ~ repwt, data = wt[-12, ]))$coefficients
```

```
             Estimate Std. Error   t value      Pr(>|t|)
(Intercept) 2.7338020 0.81479433  3.355205  9.672344e-04
repwt       0.9583743 0.01214273 78.925763 1.385270e-141
```

d. Which of the two fits in (c) would you report, and why?