Midterm Exam: Practice Set Solutions

STAT3530 Applied Linear Models

Solutions are shown in blue.

1. A lab takes $n = 5$ independent measurements of the same quantity, and a colleague proposes $\tilde{\mu} = Y_2$.

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

   $$a = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \end{bmatrix}^T$$

   since $a^T Y = 0 \cdot Y_1 + 1 \cdot Y_2 + 0 \cdot Y_3 + 0 \cdot Y_4 + 0 \cdot Y_5 = Y_2$.

   b. Is $\tilde{\mu}$ unbiased for μ?

   Yes. Every Y_i has mean μ, so $\mathbb{E}[Y] = \mu \mathbf{1}$ and

   $$\mathbb{E}[\tilde{\mu}] = a^T \mathbb{E}[Y] = \mu \, a^T \mathbf{1} = \mu \sum_i a_i = \mu$$

   because the entries of a sum to 1. Throwing away four of the five measurements costs you nothing in bias — any weights summing to one give an unbiased estimator.

   c. The least squares estimator is $\hat{\mu} = \bar{Y}$. Is it better or worse than $\tilde{\mu}$?

   Better. Both are unbiased, so compare variances:

   $$\mathrm{Var}[\tilde{\mu}] = \sigma^2 a^T a = \sigma^2 \qquad \text{against} \qquad \mathrm{Var}[\bar{Y}] = \frac{\sigma^2}{5}$$

   so the colleague's estimator is five times as variable.

   You could also answer without computing anything. $\bar{Y}$ is the least squares estimator in this model, and both estimators are linear in Y and unbiased, so Gauss-Markov says no linear unbiased estimator can beat $\bar{Y}$. The calculation only tells you *how much* worse the alternative is.

2. An instrument reads zero on a blank specimen, so $Y_i = \beta x_i + \epsilon_i$ with no intercept, and X is the single column x.

   a. Find X^TX and X^TY, writing each as a sum.

X has one column, so both are 1×1 — ordinary numbers:

$$X^TX = x^Tx = \sum_i x_i^2 \qquad \text{and} \qquad X^TY = x^TY = \sum_i x_i Y_i$$

   b. Find $(X^TX)^{-1}$ and show that $\hat{\beta} = \sum_i x_i Y_i / \sum_i x_i^2$.

Inverting a 1×1 matrix is just taking a reciprocal, so $(X^TX)^{-1} = 1/\sum_i x_i^2$ and

$$\hat{\beta} = (X^TX)^{-1}X^TY = \frac{\sum_i x_i Y_i}{\sum_i x_i^2}$$

Note this is not the usual slope formula: there are no deviations from $\bar{x}$ and $\bar{Y}$ anywhere, because with no intercept the line is not required to pass through $(\bar{x}, \bar{Y})$.

   c. Write down $\text{Var}[\hat{\beta}]$. What happens to it as the concentrations are spread further from zero?

$$\text{Var}[\hat{\beta}] = \sigma^2(X^TX)^{-1} = \frac{\sigma^2}{\sum_i x_i^2}$$

Spreading the concentrations further from zero makes $\sum_i x_i^2$ larger and the variance smaller. The estimate is a slope through the origin, so specimens far from zero are the ones that pin it down — the same reason a wider spread of predictor values helps in the model with an intercept.

   d. With $\sum_i x_i^2 = 650$, $\sum_i x_i Y_i = 1560$, $\hat{\sigma} = 0.80$, and $t_{0.975} = 2.201$, compute $\hat{\beta}$ and a 95% confidence interval.

$$\hat{\beta} = \frac{1560}{650} = 2.40, \qquad \text{se}[\hat{\beta}] = \frac{\hat{\sigma}}{\sqrt{\sum_i x_i^2}} = \frac{0.80}{\sqrt{650}} = 0.0314$$

$$2.40 \pm 2.201(0.0314) = 2.40 \pm 0.069 = (2.33, 2.47)$$

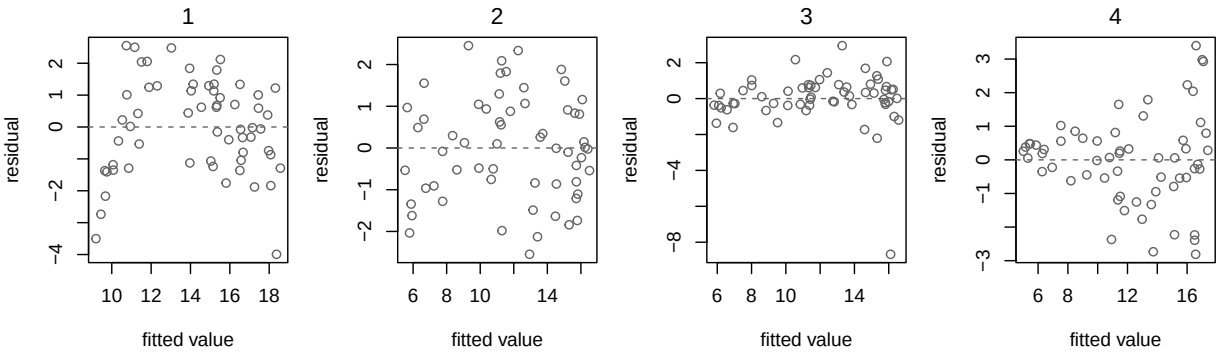
Each additional unit of concentration is estimated to raise the reading by between 2.33 and 2.47 units.

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

That the no-intercept claim is supported by the data and not just by theory. Fit the model *with* an intercept and look at it: if the intercept estimate is small and its interval comfortably contains zero, dropping it is defensible. Also check that the data include specimens at low concentrations, since with readings only at high x the fit near the origin is pure extrapolation and the constraint is doing work you cannot see. And check the residual plot as usual — forcing a line through the origin can manufacture curvature that was not there.

3. Spot checks.

   a. Name the problem each residual plot shows, or write "none".



1: nonlinearity — the residuals arc, high at the left, dipping through the middle, rising again at the right.

2: none — a shapeless band of even width.

3: one outlier, at about −6 standardized residuals, with the rest of the band clean.

4: nonconstant variance — the spread grows steadily with the fitted value.

   b. For each panel that shows a problem, propose a fix.

1: transform the *predictor*, not the response. The scatterplot rises steeply and then flattens while the spread about the curve stays the same from left to right, which is the signature of a predictor on the wrong scale. Refitting on $\log(x)$ straightens it and leaves the residual plot clean.

3: do not simply delete the point. Check whether it is influential, find out whether the record is erroneous, and if it stays, report the fit with and without it. Here it sits in the middle of the predictor range, so it is an outlier without much leverage; it inflates $\hat{\sigma}$ more than it moves the line.

4: transform the response, choosing the power with a Box-Cox plot, or model the variance explicitly. Nothing needs doing to the predictor.

Panel 2 needs no fix.

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

1: none. Curvature means the fitted line is wrong everywhere, so neither the slope nor a prediction can be trusted; nothing is salvageable without refitting.

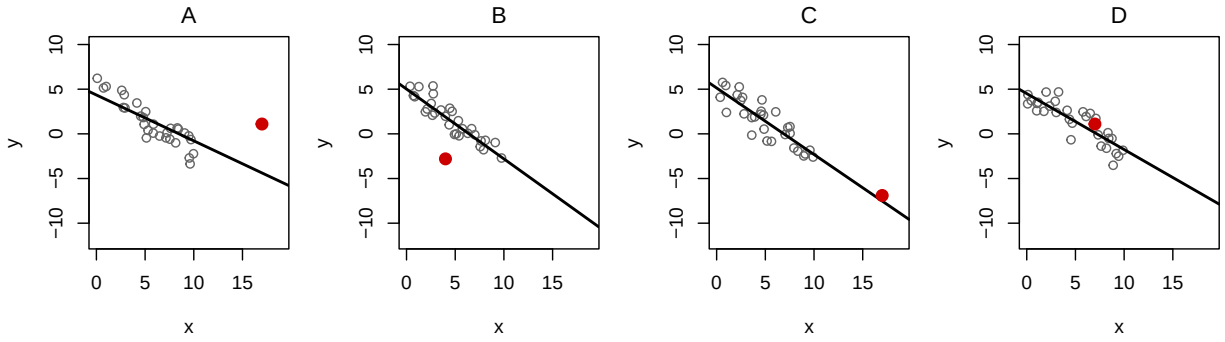
2: everything — estimation, confidence intervals, tests, and prediction intervals.

3: the estimates themselves, roughly. A single outlier at mid-range barely moves the line (here the slope shifts from 1.18 to 1.12), so the fitted line still estimates the mean response. What it damages is $\hat{\sigma}$, which it inflates, so intervals come out too wide and tests too conservative.

4: point estimates and predictions of the mean. Nonconstant variance leaves least squares unbiased, so the fitted line is still a valid estimate of the mean response. Standard errors, confidence intervals, tests, and prediction intervals are all unreliable, because they assume one common σ when the truth varies with x.

The pattern worth remembering: getting the mean function wrong (panel 1) ruins everything, while getting the error structure wrong (panels 3 and 4) leaves the fitted line usable and ruins the uncertainty statements.

d. Label each marked point as high leverage (L), an outlier (O), influential (I), or none of these (N).



A: L, O, and I. Far out in x ($h = 0.354$ against an average of 0.065) *and* far off the trend ($r = 4.25$), and the combination drags the line: the slope goes from -0.77 without it to -0.51 with it, and $D = 4.96$.

B: O. A large residual ($r = -3.79$) at a typical x ($h = 0.034$, below average), so it barely moves the line: the slope changes from -0.80 to -0.78.

C: L. Far out in x ($h = 0.396$) but sitting right on the trend ($r = 0.65$), so it is not an outlier and it is not influential; if anything it stabilizes the slope.

D: N. Ordinary in both directions ($h = 0.045$, $r = 0.84$, $D = 0.02$). It was marked to make the point that a highlighted observation is not automatically a problem.

Leverage is about x alone, outlyingness about the residual, and influence is what happens when a point has both.

4. Gestation length and body weight for 96 mammal species, with model P on the original scale and model Q on the log-log scale.

   a. Which model would you use, and why?

Model Q. Model P is badly misspecified: the species pile up at the left of the scatterplot, its residuals curve and fan out sharply with the fitted values, and its QQ plot bends at both ends. Model Q spreads the species evenly along a straight line, and its residuals are a shapeless band of roughly constant width with a QQ plot close to the line.

The transformation is doing real work here rather than being applied out of habit: gestation cannot scale linearly with body weight across species spanning grams to tonnes, and the log scale is where that relationship is straight.

   b. Report a 95% confidence interval for the slope and interpret it. What happens to gestation when a species is twice as heavy?

```
confint(fitQ)['log(Body)', ]
```

```
    2.5 %    97.5 %
0.2190595 0.2841315
```

```
2^confint(fitQ)['log(Body)', ]
```

```
   2.5 %   97.5 %
1.163975 1.217677
```

$\hat{\beta}_1 = 0.252$ with a 95% interval of $(0.219, 0.284)$. Since both variables are logged, the slope is an elasticity: a 1% increase in body weight is associated with about a 0.25% increase in mean gestation.

Doubling is easier to picture. A species twice as heavy has gestation multiplied by $2^{0.252} = 1.19$, with a 95% interval of $(1.16, 1.22)$ — about 19% longer, and somewhere between 16% and 22% longer with confidence. Gestation rises with size, but far more slowly than size itself.

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

Model Q gives the fitted value on the log scale, so exponentiate to get back to days. The appropriate interval for a single new species is the *prediction* interval:

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

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

The estimate is 174 days, with a 95% prediction interval of $(69, 442)$ days. The interval is wide because mammals of a given size vary a great deal in gestation, and it is asymmetric around the estimate because it was computed on the log scale and transformed back.

Two things worth noting. Back-transforming a fitted value from a log-scale model estimates the *median* gestation at 50 kg, not the mean. And model P's prediction interval at 50 kg runs from -30 to 309 days — a negative gestation length, which is a good sign you should not have used it.

5. Measured and self-reported weights, in kg, for 183 adults.

   a. Is the model adequately specified for inference?

No, not as it stands. Linearity is fine and the points otherwise cluster tightly along the line, but one record sits far away from every other: a measured weight of 166 kg against a reported 56 kg. It dominates the residual plot, throws the QQ plot off at one end, and is the reason $\hat{\sigma} = 8.4$ for a relationship this tight ($R^2 = 0.70$).

That observation needs dealing with before any inference is trusted, which is what parts (c) and (d) do.

   b. Are self-reports accurate on average? State the hypotheses, build the test statistic, and compare with 1.97.

Accurate on average means a one-kilogram difference in reports corresponds to a one-kilogram difference on the scale, so the claim is about the slope:

$$H_0 : \beta_1 = 1 \qquad \text{against} \qquad H_1 : \beta_1 \neq 1$$

`summary()` tests $\beta_1 = 0$, which nobody believes, so build the statistic yourself:

$$t = \frac{\hat{\beta}_1 - 1}{\text{se}[\hat{\beta}_1]} = \frac{0.9278 - 1}{0.0453} = -1.59$$

Since $|-1.59| < 1.97$, we do not reject H_0: on this fit, the data are consistent with self-reports being accurate on average.

   c. Redo the test on the fit without the flagged subject. Does your conclusion change?

```
fit.no12 <- lm(weight ~ repwt, data = wt[-12, ])
summary(fit.no12)$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
```

$$t = \frac{0.9584 - 1}{0.0121} = -3.43$$

and $|-3.43| > 1.97$, so now we reject H_0. The conclusion flips completely: without that one subject, self-reports are *not* accurate on average.

The slope estimate barely moved, from 0.928 to 0.958. What changed is its standard error, which fell from 0.0453 to 0.0121 once the wild point stopped inflating $\hat{\sigma}$ (8.42 down to 2.25, with R^2 rising from 0.70 to 0.97). A single bad record had made the estimate look four times more uncertain than it is, and that extra uncertainty was hiding a real effect.

A slope below one means reports are compressed toward the middle: heavier people under-report and lighter people over-report, by about 4% of their weight.

   d. Which of the two fits would you report, and why?

The fit without subject 12. The usual rule is not to drop observations merely for being inconvenient, but this record is demonstrably erroneous rather than merely unusual: the subject is listed as 166 kg and 57 cm tall, which is the height and weight of the same person interchanged. It is a transcription error, not a person.

Report the corrected analysis, and say plainly what you did and why — that one record was excluded as a data-entry error, and that with it included the standard error inflates enough to reverse the conclusion. A reader can then judge the decision for themselves, which is the point of disclosing it.