Homework 4 solutions

STAT3530 Applied Linear Models

Solutions are shown in blue.

Occupational prestige

1. Interpret each of the three partial significance tests.

```
fit <- lm(prestige ~ education + log(income) + women, data = prestige)
round(summary(fit)$coefficients, 4)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -110.9658    14.8429 -7.4760     0.00
education      3.7305     0.3544 10.5268     0.00
log(income)   13.4382     1.9138  7.0219     0.00
women          0.0469     0.0299  1.5685     0.12
```

i. `education` ($t = 10.53$, $p < 0.001$). Holding income and percentage women fixed, there is strong evidence that education is associated with prestige.

ii. `log(income)` ($t = 7.02$, $p < 0.001$). Holding education and percentage women fixed, there is strong evidence that income is associated with prestige. The test is about log(income), so it asks whether prestige changes with *proportional* changes in pay.

iii. `women` ($t = 1.57$, $p = 0.120$). These data do not provide evidence of an association between the percentage of women in an occupation and its prestige, once education and income are accounted for. This is a failure to reject, not evidence of no association: the interval in (2) is wide enough to contain values that would matter.

2. Report and interpret a 95% confidence interval for each coefficient.

```
round(confint(fit), 3)
```

```
               2.5 %  97.5 %
(Intercept) -140.421 -81.511
education      3.027   4.434
log(income)    9.640  17.236
women         -0.012   0.106
```

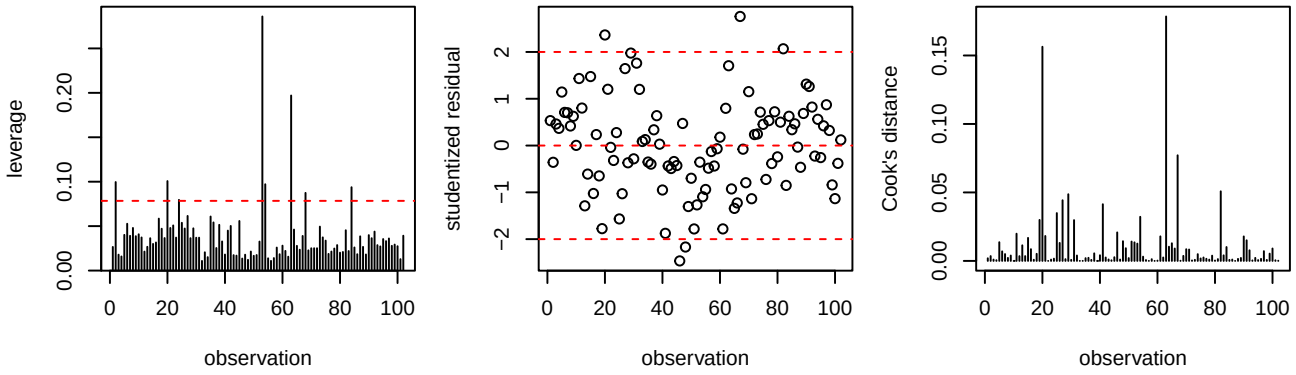
i. `education`. Each additional year of average education is associated with an increase in mean prestige of between 3.03 and 4.43 points, holding the others fixed.

ii. `log(income)`. On the log scale the interval is 9.64 to 17.24, which is not directly interpretable in dollars. Read per doubling of income, $\widehat{\beta}\log 2$ gives an increase in mean prestige of 9.31 points, with interval 6.68 to 11.95. Doubling the pay of a poorly paid occupation buys the same prestige as doubling the pay of a well paid one.

iii. `women`. The interval runs from -0.012 to 0.106 points of prestige per percentage point of women. It contains zero, and both endpoints are small: even the upper end implies only about 5 points of prestige between an all-male and an all-female occupation.

3. Influence diagnostics for `ministers` and `babysitters`.

```
par(mfrow = c(1, 3), mar = c(4, 4, 2, 1))
plot(hatvalues(fit), type = 'h', xlab = 'observation', ylab = 'leverage')
abline(h = 2 * 4 / 102, col = 'red', lty = 2)
plot(rstandard(fit), xlab = 'observation', ylab = 'studentized residual')
abline(h = c(-2, 0, 2), col = 'red', lty = 2)
plot(cooks.distance(fit), type = 'h', xlab = 'observation', ylab = "Cook's
 ↪  distance")
```




```
who <- c('ministers', 'babysitters')
k   <- match(who, prestige$occupation)

round(data.frame(h = hatvalues(fit)[k], r = rstandard(fit)[k],
                 D = cooks.distance(fit)[k], row.names = who), 3)
```

```
                h     r     D
ministers   0.101 2.363 0.156
babysitters 0.197 1.706 0.178
```

The leverage cutoff is $2p/n = 8/102 = 0.078$.

	h_{ii}	r_i	D_i	high leverage?	outlier?
`ministers`	0.101	2.363	0.156	yes, mildly	**yes**
`babysitters`	0.197	1.706	0.178	**yes**	no

They reach almost the same Cook's distance from opposite directions.

4. Why each is unusual.

```
prestige[k, ]
```

```
    occupation prestige education income women
20   ministers     72.8     14.50   4686  4.14
63 babysitters     25.9      9.46    611 96.53
```

i. `ministers` are an outlier in the response. With 14.5 years of education but an income of only \$4,686, the model predicts a prestige of 56.9; the actual rating was 72.8. Clergy are respected far beyond what their pay would suggest, and no combination of these three predictors captures that. Their leverage is only slightly above the cutoff, so it is the residual doing the work.

ii. `babysitters` are extreme in the predictors. Their income of \$611 is the lowest of all 102 occupations, and at 96.5% women they are near the top of that range as well — a corner of predictor space with almost no other occupations in it. Their residual is unremarkable ($r = 1.71$), so they are not an outlier; they simply sit far from the centre of the data, which is what leverage measures.

5. Refits.

```
drop.one <- function(who) {
  m <- lm(prestige ~ education + log(income) + women,
          data = prestige[!prestige$occupation %in% who, ])
  summary(m)$coefficients['women', c('Estimate', 'Pr(>|t|)')]
}

round(rbind(all.102        = summary(fit)$coefficients['women', c('Estimate',
 ↪  'Pr(>|t|)')],
            no.ministers   = drop.one('ministers'),
            no.babysitters = drop.one('babysitters'),
            neither        = drop.one(c('ministers', 'babysitters'))), 4)
```

```
               Estimate Pr(>|t|)
all.102          0.0469   0.1200
no.ministers     0.0643   0.0349
no.babysitters   0.0519   0.0842
neither          0.0714   0.0186
```

6. Does the inference change? Is it sensitive to one occupation or both?

Yes, it changes. With all 102 occupations the `women` coefficient is 0.047 with $p = 0.120$ and we do not reject; dropping both occupations gives 0.071 with $p = 0.019$, which we would report as a significant positive association.

It is mostly `ministers`. Dropping them alone takes p from 0.120 to 0.035, which crosses the usual threshold on its own. Dropping `babysitters` alone takes p only to 0.084, which does not. Dropping both moves it furthest because they push in the same direction: `ministers` is a high-prestige, low-income, overwhelmingly male occupation, and `babysitters` is a low-prestige, low-income, overwhelmingly female one, so each on its own drags the fitted relationship between `women` and prestige toward zero.

The honest summary is that a conclusion that depends on 1 or 2 observations out of 102 is not a conclusion.

7. What would you report?

That these data do not settle it. The point estimate is positive under every version of the fit, but whether it is distinguishable from zero depends on whether two particular occupations are included, and there is no statistical reason to exclude them — neither is a data error, and both are genuine occupations in the population of interest.

The right report gives the interval from the full data, states that two influential occupations were identified, shows what happens without them, and says the evidence is suggestive but not conclusive. Reporting only the fit that reaches $p < 0.05$ would be choosing the answer first.

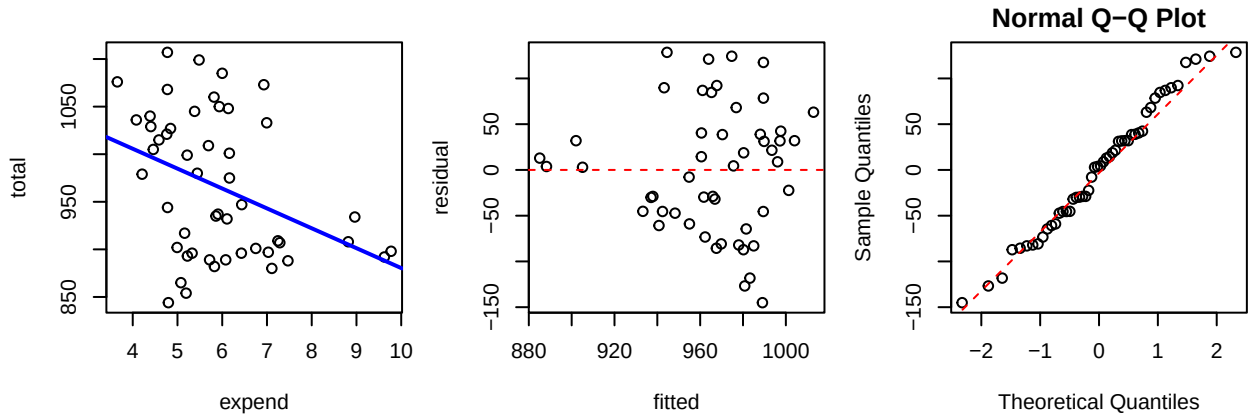
What explains SAT scores

8. Regress `total` on `expend`.

```
a <- lm(total ~ expend, data = sat)
round(summary(a)$coefficients, 3)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1089.294     44.390  24.539    0.000
expend       -20.892      7.328  -2.851    0.006
```

```
par(mfrow = c(1, 3), mar = c(4, 4, 2, 1))
plot(sat$expend, sat$total, xlab = 'expend', ylab = 'total'); abline(a, col =
 ↪  'blue', lwd = 2)
plot(fitted(a), resid(a), xlab = 'fitted', ylab = 'residual'); abline(h = 0, col =
 ↪  'red', lty = 2)
qqnorm(resid(a)); qqline(resid(a), col = 'red', lty = 2)
```



Nothing needs transforming: the scatterplot is straight, the residuals are shapeless, and the QQ plot is fine. The slope is -20.89 with $p = 0.006$, so states that spend an additional thousand dollars per pupil average about 21 points *lower*. Taken alone, this fit appears to support the legislator.

9. Adjust for `takers`.

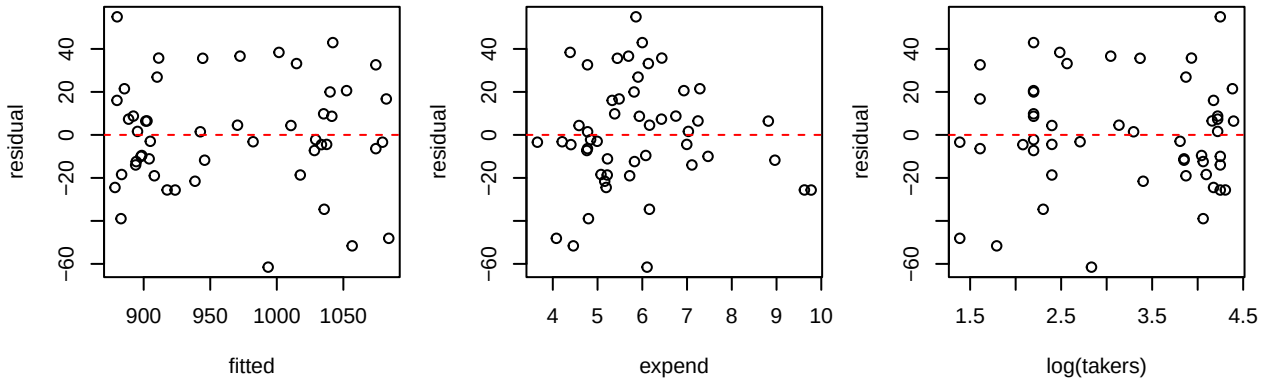
Entered untransformed, `takers` is misspecified — a quadratic term in it is significant at $p = 5 \times 10^{-6}$, and the residual plot bends. On the log scale it is straight:

```
b <- lm(total ~ expend + log(takers), data = sat)
round(summary(b)$coefficients, 3)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1147.113     16.700  68.688    0.000
expend        11.130      3.264   3.410    0.001
log(takers)  -78.205      4.471 -17.491    0.000
```

```
par(mfrow = c(1, 3), mar = c(4, 4, 2, 1))
plot(fitted(b), resid(b), xlab = 'fitted', ylab = 'residual'); abline(h = 0, col =
 ↪  'red', lty = 2)
plot(sat$expend, resid(b), xlab = 'expend', ylab = 'residual'); abline(h = 0, col =
 ↪  'red', lty = 2)
```

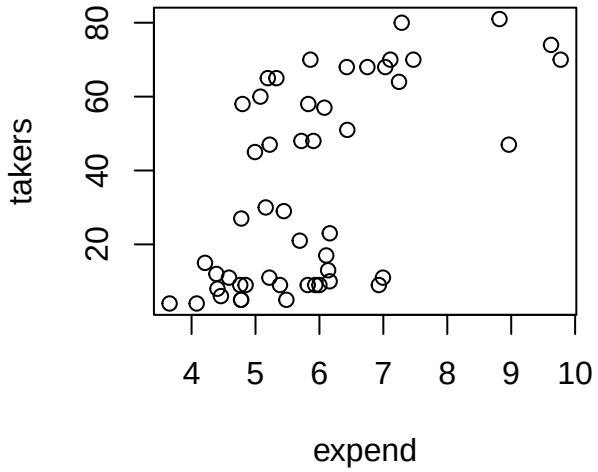
```
plot(log(sat$takers), resid(b), xlab = 'log(takers)', ylab = 'residual'); abline(h =
↪  0, col = 'red', lty = 2)
```



Adjusted for participation, the spending coefficient is +11.13 with $p = 0.001$: comparing two states where the same share of students sits the exam, the one spending an additional thousand dollars per pupil averages about 11 points **higher**. The sign has flipped and it is still significant.

10. Explain the discrepancy.

```
par(mar = c(4, 4, 1, 1))
plot(sat$expend, sat$takers, xlab = 'expend', ylab = 'takers')
```




```
round(cor(sat$expend, sat$takers), 3)
```

```
[1] 0.593
```

The two predictors are strongly related ($r = 0.59$): high-spending states are mostly states where a large share of students takes the SAT. Participation is what drives the score down — where nearly everyone sits the exam, the average includes weaker students, while in low-participation states only the college-bound self-select into it.

So the simple fit in (8) compares states that differ in spending *and* in participation at the same time, and attributes the participation effect to spending. Once `takers` is held fixed, the confounding is removed and the remaining association with spending is positive.

11. Predictions at fixed participation.

```
new <- data.frame(expend = c(6, 7), takers = c(30, 30))

round(predict(b, new, interval = 'confidence'), 1)
```

```
    fit   lwr   upr
1 947.9 940.3 955.5
2 959.0 949.4 968.7
```

```
round(predict(b, new, interval = 'prediction'), 1)
```

```
    fit   lwr    upr
1 947.9 895.5 1000.3
2 959.0 906.3 1011.8
```

i. At \$6,000 per pupil with 30% taking the exam, the estimated mean total is 947.9, with a 95% confidence interval of (940.3, 955.5) for the mean and a prediction interval of (895.5, 1000.3) for a single state. At \$7,000 with the same participation, the estimated mean is 959.0, with intervals (949.4, 968.7) and (906.3, 1011.8).

ii. The difference is $959.0 - 947.9 = 11.1$, exactly the coefficient on `expend`. That is what the coefficient *means*: the two rows of `new` differ by one unit of `expend` and nothing else, so everything contributed through log(takers) and the intercept cancels, leaving $\hat{\beta}_{\text{expend}} \times 1$. Holding a predictor fixed is not a figure of speech here — it is the arithmetic.

Note how much wider the prediction intervals are. The model locates the *mean* for such states to within about 8 points, but individual states scatter around that mean by far more.

12. Root mean square prediction error for the model in (9).

```
rmspe <- function(m) sqrt(mean((resid(m) / (1 - hatvalues(m)))^2))

round(rmspe(b), 2)
```

```
[1] 26.57
```

For a state that was not used to fit the model, this model's prediction of the average SAT total would be off by roughly 27 points. The state averages themselves run from 844 to 1107, so that is a useful but not precise prediction.

13. Add `salary` and `ratio`.

```
e <- lm(total ~ expend + log(takers) + salary + ratio, data = sat)

round(c(two.predictors = rmspe(b), four.predictors = rmspe(e)), 2)
```

```
 two.predictors four.predictors
          26.57           27.16
```

No. Predictions get *worse*, from 26.57 to 27.16. Neither new predictor is close to significant, and each costs a coefficient estimated from the same 50 states, so the extra flexibility buys noise rather than signal.

14. Compare in sample instead.

```
in.sample <- function(m) mean(resid(m)^2)

round(rbind(mean.sq.resid = c(two = in.sample(b), four = in.sample(e)),
            leave.one.out = c(mean((resid(b)/(1 - hatvalues(b)))^2),
                              mean((resid(e)/(1 - hatvalues(e)))^2))), 1)
```

```
                two  four
mean.sq.resid 624.8 612.4
leave.one.out 706.0 737.9
```

No, it is the opposite story. In sample the four-predictor model looks better (612.4 against 624.8); out of sample it is worse (737.9 against 706.0).

The leave-one-out comparison is the one to trust. The in-sample average is computed from the same observations that chose the coefficients, so adding any column — even pure noise — can only reduce it. It therefore cannot distinguish a model that has learned something from one that has fit the sample better by accident, which is exactly the comparison being asked for here.