Homework 6 solutions

STAT3530 Applied Linear Models

Solutions are shown in blue.

Educational expenditure and SAT scores (again)

1. Fitted model equation.

```
fit <- lm(total ~ expend + takers + I(takers^2), data = sat)
round(summary(fit)$coefficients, 4)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1051.8871    20.8248 50.5113   0.0000
expend         7.9137     3.4982  2.2622   0.0285
takers        -6.3807     0.7036 -9.0684   0.0000
I(takers^2)    0.0474     0.0092  5.1752   0.0000
```

$$\widehat{\texttt{total}} = 1051.89 + 7.91\,\texttt{expend} - 6.381\,\texttt{takers} + 0.0474\,\texttt{takers}^2$$

Use `I(takers^2)`, not `poly(takers, 2)`: same fit, but (3)–(5) need the raw coefficients.

2. Test for curvature.

$$H_0 : \beta_3 = 0 \qquad \text{against} \qquad H_1 : \beta_3 \neq 0$$

```
anova(lm(total ~ expend + takers, data = sat), fit)
```

```
Analysis of Variance Table

Model 1: total ~ expend + takers
Model 2: total ~ expend + takers + I(takers^2)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     47 49520
2     46 31298  1     18222 26.783 4.872e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$F_{1,46} = 26.78$, $p < 0.001$. Strong evidence of curvature: mean total score is not a straight-line function of participation. Equivalently $t = 5.175$ from (1), and $t^2 = F$.

3. Rate of change.

$$\frac{\partial}{\partial \texttt{takers}} E(\texttt{total}) = \beta_2 + 2\beta_3\,\texttt{takers} \qquad \text{estimated by} \qquad -6.381 + 0.0948\,\texttt{takers}$$

```
b <- coef(fit)
slope <- function(tk) b['takers'] + 2 * b['I(takers^2)'] * tk

round(slope(c(10, 40, 70)), 3)
```

```
[1] -5.432 -2.588  0.257
```

participation	rate of change of $E(\texttt{total})$
10%	−5.43 points per percentage point
40%	−2.59 points per percentage point
70%	+0.26 points per percentage point

Participation costs about 5.4 points per percentage point where few students test, 2.6 points at 40%, and nothing at 70%.

4. Turning point.

Set the derivative to zero: $\beta_2 + 2\beta_3\theta = 0$, so

$$\theta = -\frac{\beta_2}{2\beta_3}$$

```
theta <- function(f) { b <- coef(f); unname(-b[3] / (2 * b[4])) }

theta(fit)
```

```
[1] 67.29094
```

$\hat{\theta} = 67.29\%$. It is a ratio of coefficients, not a contrast, so `summary()` gives no standard error for it — hence (5).

5. Simulated sampling distribution and percentile interval.

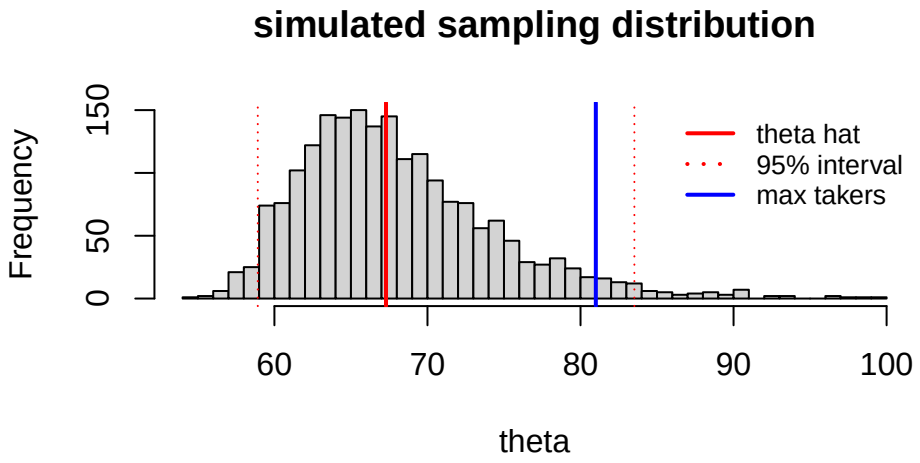
```
refit <- function(y) lm(y ~ expend + takers + I(takers^2), data = sat)

set.seed(5)
theta.sim <- replicate(2000, {
  y <- fitted(fit) + sample(resid(fit))
  refit(y) |> theta()
})

round(quantile(theta.sim, c(0.025, 0.975)), 2)
```

```
 2.5% 97.5%
58.91 83.52
```

```
hist(theta.sim, breaks = 40, xlab = 'theta', main = 'simulated sampling
 ↪  distribution')
abline(v = theta(fit), col = 'red', lwd = 2)
abline(v = quantile(theta.sim, c(0.025, 0.975)), col = 'red', lty = 3)
abline(v = max(sat$takers), col = 'blue', lwd = 2)
legend('topright', bty = 'n', cex = 0.8,
       legend = c('theta hat', '95% interval', 'max takers'),
       col = c('red', 'red', 'blue'), lty = c(1, 3, 1), lwd = 2)
```



We are 95% confident the participation rate at which mean total SAT score stops falling is between 59% and 84%. The distribution is right skewed because $\hat{\theta}$ divides by $2\hat{\beta}_3$.

6. Is the turning point real?

```
c(upper = unname(quantile(theta.sim, 0.975)), max.takers = max(sat$takers))
```

```
    upper max.takers
 83.51993   81.00000
```

No. The upper endpoint 83.5 runs past the largest observed participation rate, 81, so the data cannot locate the turn inside their own range — they are compatible with the curve still falling everywhere we can see it. Only 10 states lie above $\hat{\theta}$.

A parabola must turn somewhere once curvature is admitted, so the turn is a property of the specification. The log model from HW4 bends the same way, never turns, and fits marginally better:

```
fit.log <- lm(total ~ expend + log(takers), data = sat)

round(c(quadratic = sigma(fit), log = sigma(fit.log)), 2)
```

```
quadratic       log
    26.08     25.78
```

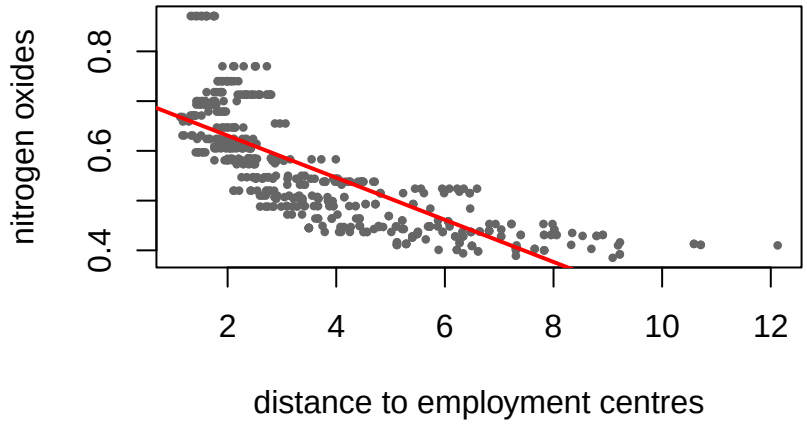
Air pollution near Boston

7. Simple linear regression of `nox` on `dis`.

```
m.lin <- lm(nox ~ dis, data = boston)
round(summary(m.lin)$coefficients, 4)
```

```
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)   0.7153     0.0068 105.2565        0
dis          -0.0423     0.0016 -27.0265        0
```

```
plot(nox ~ dis, data = boston, col = 'grey40', pch = 16, cex = 0.6,
     xlab = 'distance to employment centres', ylab = 'nitrogen oxides')
abline(m.lin, col = 'red', lwd = 2)
```



Strongly negative and significant ($t = -27.0$, $p < 0.001$, $R^2 = 0.59$): tracts further from the employment centres have cleaner air, by about 0.042 parts per 10 million per unit distance.

Not well specified. The relationship is convex, not straight — pollution drops steeply over the first few units and then levels off, so the line sits below the cloud at both ends and above it in the middle:

```
round(tapply(resid(m.lin), cut(boston$dis, c(1, 2, 3, 4, 6, 13)), mean), 3)
```

```
 (1,2]  (2,3]  (3,4]  (4,6] (6,13]
 0.048 -0.001 -0.057 -0.036  0.032
```

Binned residual means that change sign with x are the signature of a misspecified mean function.

8. Piecewise linear, continuous, knot at $t = 3$.

   i. Model.

$$\texttt{nox}_i = \gamma_0 + \gamma_1 \texttt{dis}_i + \gamma_2 (\texttt{dis}_i - 3)_+ + \epsilon_i, \qquad (x - t)_+ = \max(x - t, 0)$$

   The hinge term is flat below the knot, so γ_2 is the *change* in slope at 3.

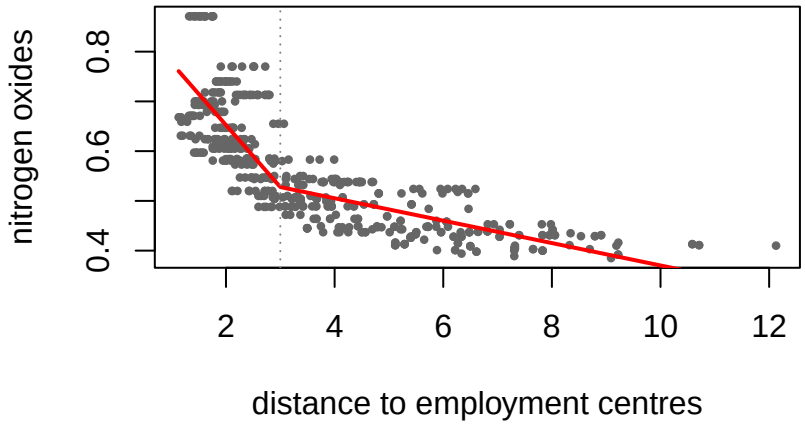
   ii. Fitted model equation.

```
m.pw <- lm(nox ~ dis + pmax(dis - 3, 0), data = boston)
round(summary(m.pw)$coefficients, 4)
```

```
                Estimate Std. Error  t value Pr(>|t|)
(Intercept)       0.9016     0.0143  63.1684        0
dis              -0.1246     0.0059 -21.0410        0
pmax(dis - 3, 0)  0.1021     0.0072  14.2519        0
```

$$\widehat{\texttt{nox}} = 0.9016 - 0.1246\,\texttt{dis} + 0.1021\,(\texttt{dis} - 3)_+$$

```
grid <- data.frame(dis = seq(min(boston$dis), max(boston$dis), length = 300))
plot(nox ~ dis, data = boston, col = 'grey40', pch = 16, cex = 0.6,
     xlab = 'distance to employment centres', ylab = 'nitrogen oxides')
lines(grid$dis, predict(m.pw, grid), col = 'red', lwd = 2)
abline(v = 3, col = 'grey50', lty = 3)
```



iii. Slopes either side of the knot.

```
g <- coef(m.pw)

round(c(below = unname(g[2]), above = unname(g[2] + g[3])), 4)
```

```
  below   above
-0.1246 -0.0225
```

$$\texttt{dis} \le 3: \quad \hat{\gamma}_1 = -0.1246 \qquad \texttt{dis} > 3: \quad \hat{\gamma}_1 + \hat{\gamma}_2 = -0.0225$$

Within three units of the centres each additional unit of distance buys 0.125 fewer parts per 10 million; beyond three units, only 0.023.

9. Does the bend explain more variation?

$$H_0 : \gamma_2 = 0 \qquad \text{against} \qquad H_1 : \gamma_2 \neq 0$$

```
anova(m.lin, m.pw)
```

```
Analysis of Variance Table

Model 1: nox ~ dis
Model 2: nox ~ dis + pmax(dis - 3, 0)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    504 2.7686
2    503 1.9722  1   0.79639 203.12 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$F_{1,503} = 203.1$, $p < 0.001$. Overwhelming evidence that the bend at 3 explains air pollution the straight line does not: the rate at which air quality improves with distance differs near the employment centres and far from them.

10. Degree 1 b-spline basis.

    i. Knots and basis functions.

    One interior knot, $t_1 = 3$, plus the two boundary knots at the extremes of `dis`, $t_0 = 1.13$ and $t_2 = 12.13$ — three knots, and so **three basis functions** h_0, h_1, h_2. In general $1 + K + 1$ columns for degree 1 with K interior knots and an intercept; the same three dimensions as (8).

    ii. Model and estimates.

$$\texttt{nox}_i = c_0 h_0(\texttt{dis}_i) + c_1 h_1(\texttt{dis}_i) + c_2 h_2(\texttt{dis}_i) + \epsilon_i$$

```
m.bs <- lm(nox ~ bs(dis, knots = 3, degree = 1, intercept = TRUE) - 1,
           data = boston)

round(coef(m.bs), 4) |> setNames(c('c0', 'c1', 'c2'))
```

```
    c0     c1     c2
0.7608 0.5278 0.3223
```

$$\hat{c}_0 = 0.7608 \qquad \hat{c}_1 = 0.5278 \qquad \hat{c}_2 = 0.3223$$

Each h_k is zero at every knot but t_k, so $c_k = f(t_k)$: the coefficients are the fitted mean `nox` at the three knots.

    iii. Recover the slopes.

$$\text{slope on } [t_{k-1}, t_k] = \frac{c_k - c_{k-1}}{t_k - t_{k-1}}$$

```
cc <- coef(m.bs)
kn <- c(min(boston$dis), 3, max(boston$dis))

round(diff(cc) / diff(kn), 4) |> setNames(c('below', 'above'))
```

```
  below   above
-0.1246 -0.0225
```

$$\text{below: } \frac{0.5278 - 0.7608}{3 - 1.13} = -0.1246 \qquad \text{above: } \frac{0.3223 - 0.5278}{12.13 - 3} = -0.0225$$

Matches 8(iii). Both bases span the same space, so the fit is identical — only the coordinates differ.

```
c(rss.hinge = deviance(m.pw), rss.spline = deviance(m.bs))
```

```
 rss.hinge rss.spline
  1.972174   1.972174
```