Homework 5 solutions

STAT3530 Applied Linear Models

Solutions are shown in blue.

Seat position

```
full <- lm(hipcenter ~ ., data = seatpos)
red  <- lm(hipcenter ~ Age + Ht + Thigh, data = seatpos)
```

1. Test whether the five dropped predictors matter.

$$H_0 : \beta_{\texttt{Weight}} = \beta_{\texttt{HtShoes}} = \beta_{\texttt{Seated}} = \beta_{\texttt{Arm}} = \beta_{\texttt{Leg}} = 0 \qquad \text{against} \qquad H_1 : \text{at least one is nonzero}$$

Five constraints, so 5 numerator df; the denominator is the residual df of the full model, $38 - 9 = 29$.

```
anova(red, full)
```

```
Analysis of Variance Table

Model 1: hipcenter ~ Age + Ht + Thigh
Model 2: hipcenter ~ Age + Weight + HtShoes + Ht + Seated + Arm + Thigh +
    Leg
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     34 45070
2     29 41262  5    3808.7 0.5354 0.7477
```

$F_{5,29} = 0.54$, $p = 0.748$. No evidence that `Weight`, `HtShoes`, `Seated`, `Arm` and `Leg` explain any seat position left unexplained by age, height and thigh circumference.

2. Partial inference on `Ht` in the full and reduced models.

```
round(rbind(full = summary(full)$coefficients['Ht', ],
            reduced = summary(red)$coefficients['Ht', ]), 4)
```

```
        Estimate Std. Error t value Pr(>|t|)
full      0.6013    10.1299  0.0594   0.9531
reduced  -3.9688     0.8123 -4.8858   0.0000
```

Full model ($t = 0.06$, $p = 0.953$): no evidence of an association between height and hip position with the other seven measurements held fixed. **Reduced model** ($t = -4.89$, $p < 0.001$): holding age and thigh circumference fixed, each additional cm of height is associated with the hips about 4.0 mm further back.

The tests ask different questions — each asks what height adds *given the rest of that particular model*, and in the full model the rest includes `HtShoes`.

```
round(c(HtShoes = cor(seatpos$Ht, seatpos$HtShoes),
        Leg     = cor(seatpos$Ht, seatpos$Leg),
        Thigh   = cor(seatpos$Ht, seatpos$Thigh)), 3)
```

```
HtShoes      Leg   Thigh
  0.998    0.910   0.735
```

Height with and without shoes correlate 0.998, so the full model asks about a comparison the data barely contain. The standard error falls from 10.13 to 0.81 when the other length measures are dropped.

3. Do `Age` and `Thigh` add anything to a model with `Ht` alone?

$$H_0 : \beta_{\texttt{Age}} = \beta_{\texttt{Thigh}} = 0 \qquad \text{against} \qquad H_1 : \text{at least one is nonzero}$$

Two constraints, so 2 numerator df and $38 - 4 = 34$ denominator df.

```
ht <- lm(hipcenter ~ Ht, data = seatpos)

anova(ht, red)
```

```
Analysis of Variance Table

Model 1: hipcenter ~ Ht
Model 2: hipcenter ~ Age + Ht + Thigh
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     36 47616
2     34 45070  2    2545.3 0.9601  0.393
```

$F_{2,34} = 0.96$, $p = 0.393$. No evidence that age and thigh circumference explain anything beyond height. Two group tests in a row have taken us from eight predictors to one without objection.

4. Root mean square prediction error.

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

round(c(full = rmspe(full), reduced = rmspe(red), Ht.only = rmspe(ht)), 2)
```

```
   full reduced Ht.only
  44.45   38.47   37.25
```

model	RMSPE
full, all eight predictors	44.45 mm
reduced, `Age + Ht + Thigh`	38.47 mm
`Ht` alone	37.25 mm

No — prediction gets steadily *worse* as measurements are added. Every extra predictor costs a coefficient estimated from the same 38 drivers and carries almost no information `Ht` does not already carry. For scale, `hipcenter` has standard deviation 59.7 mm.

5. Does it matter *which* measurement you collect?

```
thigh <- lm(hipcenter ~ Thigh, data = seatpos)

round(c(Ht.only = rmspe(ht), Thigh.only = rmspe(thigh)), 2)
```

```
   Ht.only Thigh.only
     37.25      50.15
```

```
round(c(Ht.only = summary(ht)$r.squared, Thigh.only = summary(thigh)$r.squared), 3)
```

```
   Ht.only Thigh.only
     0.638      0.350
```

Yes. Thigh circumference alone predicts to within 50 mm against 37 mm for height, and explains 35% of the variation against 64%. So *how many* measurements you collect barely matters past the first, but *which* one does.

6. Recommendation for the engineer.

Measure standing height without shoes, and stop there. Expect to place the hip point to within roughly 37 mm for a driver not in the sample.

The intern is right that body measurements explain seat position and wrong about what follows. If several measurements carry the same information, that is a reason to collect one of them, not all of them. Five of the eight predictors drop out at $p = 0.75$ (1), age and thigh circumference drop on top of that at $p = 0.39$ (3), and prediction error improves from 44 mm to 37 mm as they go (4).

Two caveats: which single measurement you pick does matter (5), so this is a recommendation for height specifically; and with 38 drivers the gap between the one- and three-predictor models is small enough to be noise. The solid finding is that eight predictors are worse than one.

Savings rates

```
sv <- data.frame(sr = savings$sr,
                 scale(savings[, c('pop15', 'pop75', 'dpi', 'ddpi')]))

fit <- lm(sr ~ ., data = sv)
round(summary(fit)$coefficients, 4)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6710     0.5378 17.9832   0.0000
pop15        -4.2207     1.3237 -3.1885   0.0026
pop75        -2.1833     1.3987 -1.5610   0.1255
dpi          -0.3338     0.9226 -0.3618   0.7192
ddpi          1.1758     0.5631  2.0882   0.0425
```

Each coefficient is the change in mean savings rate, in percentage points, per one standard deviation increase in that predictor. Every test below imposes one constraint, so each has 1 and $50 - 5 = 45$ degrees of freedom.

7. Do the two demographic variables matter equally?

$$H_0 : \beta_{15} = \beta_{75} \qquad \text{against} \qquad H_1 : \beta_{15} \neq \beta_{75}, \qquad c^T = (0,\ 1,\ -1,\ 0,\ 0)$$

```
car::linearHypothesis(fit, 'pop15 = pop75')
```

```
Linear hypothesis test:
pop15 - pop75 = 0

Model 1: restricted model
Model 2: sr ~ pop15 + pop75 + dpi + ddpi

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     46 719.35
2     45 650.71  1    68.638 4.7467 0.03463 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$F_{1,45} = 4.75$, $p = 0.035$. Reject: the two age-structure variables do not matter equally. A one standard deviation increase in the share under 15 depresses the savings rate about 2.0 points more than the same shift in the share over 75.

Note β_{75} is not individually significant ($p = 0.126$) while the *difference* is; neither claim implies the other.

8. Do the two economic variables matter equally?

$$H_0 : \beta_{\texttt{dpi}} = \beta_{\texttt{ddpi}} \qquad \text{against} \qquad H_1 : \beta_{\texttt{dpi}} \neq \beta_{\texttt{ddpi}}, \qquad c^T = (0,\ 0,\ 0,\ 1,\ -1)$$

```
car::linearHypothesis(fit, 'dpi = ddpi')
```

```
Linear hypothesis test:
dpi - ddpi = 0

Model 1: restricted model
Model 2: sr ~ pop15 + pop75 + dpi + ddpi

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     46 687.21
2     45 650.71  1    36.501 2.5242 0.1191
```

$F_{1,45} = 2.52$, $p = 0.119$. Fail to reject: no evidence that the level of income and its growth rate have different effects. The estimated difference is -1.51 points and not small, so this is a failure to reject, not a demonstration of equality.

9. Do age structure and the economy matter equally?

$$H_0 : \beta_{15} + \beta_{75} = \beta_{\texttt{dpi}} + \beta_{\texttt{ddpi}} \qquad \text{against} \qquad H_1 : \text{not equal}, \qquad c^T = (0,\ 1,\ 1,\ -1,\ -1)$$

```
car::linearHypothesis(fit, 'pop15 + pop75 = dpi + ddpi')
```

```
Linear hypothesis test:
pop15  + pop75 - dpi - ddpi = 0

Model 1: restricted model
Model 2: sr ~ pop15 + pop75 + dpi + ddpi

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     46 740.91
2     45 650.71  1    90.196 6.2375 0.01623 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$F_{1,45} = 6.24$, $p = 0.016$. Reject: evidence that the two demographic variables together and the two economic variables together do not have the same combined effect.

This is one constraint, not two: the hypothesis says only that the two sums agree.

10. Point estimate for the contrast in (9).

```
b <- coef(fit); cc <- c(0, 1, 1, -1, -1)

est <- sum(cc * b)
se  <- sqrt(drop(t(cc) %*% vcov(fit) %*% cc))

round(c(estimate = est, se = se,
        lwr = est - qt(0.975, 45) * se,
        upr = est + qt(0.975, 45) * se), 3)
```

```
estimate       se      lwr      upr
  -7.246    2.901  -13.090   -1.402
```

$c^T\hat{\beta} = -7.25$ percentage points, 95% interval $(-13.09, -1.40)$.

Raising both demographic variables by one standard deviation changes the mean savings rate by -6.40 points; raising both economic variables by one standard deviation changes it by $+0.84$. The contrast is the gap: age structure moves the savings rate about 7.2 points further down than the economy moves it up. The interval is wide, so the direction is clear but the magnitude is not.

Prostate cancer

```
fit <- lm(lpsa ~ ., data = prostate)
round(summary(fit)$coefficients, 4)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.6693     1.2964  0.5163   0.6069
lcavol         0.5870     0.0879  6.6767   0.0000
lweight        0.4545     0.1700  2.6731   0.0090
age           -0.0196     0.0112 -1.7576   0.0823
lbph           0.1071     0.0584  1.8316   0.0704
svi            0.7662     0.2443  3.1360   0.0023
lcp           -0.1055     0.0910 -1.1589   0.2496
gleason        0.0451     0.1575  0.2867   0.7750
pgg45          0.0045     0.0044  1.0235   0.3089
```

11. Do any of the measurements explain log PSA?

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_8 = 0 \qquad \text{against} \qquad H_1 : \text{at least one is nonzero}$$

The omnibus test: the reduced model is intercept-only, so 8 numerator and $97 - 9 = 88$ denominator df.

```
round(summary(fit)$fstatistic, 2)
```

```
value numdf dendf
20.86  8.00 88.00
```

$F_{8,88} = 20.86$, $p < 0.001$, $R^2 = 0.65$. Overwhelming evidence that the eight measurements taken together explain log PSA. This says nothing about which ones.

12. Do the three patient variables explain log PSA?

    i. Hypotheses, degrees of freedom, result.

$$H_0 : \beta_{\texttt{lweight}} = \beta_{\texttt{age}} = \beta_{\texttt{lbph}} = 0 \qquad \text{against} \qquad H_1 : \text{at least one is nonzero}$$

Three constraints, so 3 numerator and 88 denominator df; the reduced model keeps the five tumor variables.

```
tumor <- lm(lpsa ~ lcavol + lcp + svi + gleason + pgg45, data = prostate)

anova(tumor, fit)
```

```
Analysis of Variance Table

Model 1: lpsa ~ lcavol + lcp + svi + gleason + pgg45
Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     91 52.689
2     88 44.163  3    8.5257 5.6628 0.001357 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$F_{3,88} = 5.66$, $p = 0.0014$. Reject: the patient measurements explain variation in log PSA that the tumor measurements do not. Clinically, a raised PSA need not mean a larger or more aggressive cancer.

ii. Compare with inference by partial t tests.

```
round(summary(fit)$coefficients[c('lweight', 'age', 'lbph'), ], 4)
```

```
        Estimate Std. Error t value Pr(>|t|)
lweight   0.4545     0.1700  2.6731   0.0090
age      -0.0196     0.0112 -1.7576   0.0823
lbph      0.1071     0.0584  1.8316   0.0704
```

Not all individually significant: only `lweight` reaches the usual threshold ($p = 0.009$); `age` ($p = 0.082$) and `lbph` ($p = 0.070$) do not.

Not a discrepancy. The F test asks whether the three *together* add anything to the tumor-only model; each t test asks whether one adds anything given the other seven, including the other two patient variables. Those three correlate 0.31 to 0.44, so the F test credits the group with their whole joint contribution while each t test credits a variable only with what its companions leave over.