# LING82100: homework 5 solution

## 1 Multiple regression

```
d <- read.table(
        "http://www.wellformedness.com/courses/LING82100/Data/ANAE-O.tsv",
        header = TRUE
)
```

1. ```
   > euclidean.distance <- function(x1, y1, x2, y2) {
   +     x.delta <- x1 - x2
   +     y.delta <- y1 - y2
   +     return(sqrt(x.delta * x.delta + y.delta * y.delta))
   + }
   > d$distance <- with(d, euclidean.distance(o.F1, o.F2, oh.F1, oh.F2))
   ```

2. ```
   > d$age <- d$age - mean(d$age, na.rm = TRUE)
   > contrasts(d$gender) <- contr.sum
   > contrasts(d$dialect) <- contr.sum
   > r <- lm(distance ~ age + gender + dialect, data = d)
   ```

3. ```
   > summary(r)

   Call:
   lm(formula = distance ~ age + gender + dialect, data = d)

   Residuals:
       Min      1Q  Median      3Q     Max
   -251.87  -49.96   -3.71   40.43  364.24

   Coefficients:
               Estimate Std. Error t value Pr(>|t|)
   (Intercept)  176.8756     7.8732  22.465  < 2e-16 ***
   age            0.7735     0.2954   2.618 0.009177 **
   gender1        4.4896     4.4425   1.011 0.312821
   dialect1    -103.6565    30.3734  -3.413 0.000708 ***
   dialect2    -138.7272    34.7547  -3.992 7.80e-05 ***
   dialect3    -132.0029    18.7318  -7.047 8.00e-12 ***
   ```
   ```

```
dialect4      12.7975    48.6945   0.263 0.792830
dialect5       1.5156    48.6868   0.031 0.975181
dialect6    -137.7296    59.3633  -2.320 0.020835 *
dialect7      41.2856    37.9317   1.088 0.277062
dialect8     154.6203    13.2496  11.670  < 2e-16 ***
dialect9    -100.3457    26.1983  -3.830 0.000148 ***
dialect10    155.5976    23.4500   6.635 1.05e-10 ***
dialect11    -19.0749    13.2427  -1.440 0.150530
dialect12    401.9196    38.0435  10.565  < 2e-16 ***
dialect13     31.2170    15.4657   2.018 0.044207 *
dialect14   -123.2709    32.2997  -3.816 0.000157 ***
dialect15    207.0647    48.5053   4.269 2.45e-05 ***
dialect16     80.7187    28.7859   2.804 0.005290 **
dialect17    -28.9923    12.9599  -2.237 0.025829 *
dialect18   -110.8746    83.3419  -1.330 0.184155
dialect19    -96.4842    30.3057  -3.184 0.001567 **
dialect20     19.9923    17.9933   1.111 0.267189
dialect21    -97.0180    14.3511  -6.760 4.86e-11 ***
dialect22     59.3905    24.3529   2.439 0.015170 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 86.8 on 402 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6019,Adjusted R-squared:  0.5781
F-statistic: 25.33 on 24 and 402 DF,  p-value: < 2.2e-16

> drop1(r, test = "Chisq")
Single term deletions

Model:
distance ~ age + gender + dialect
        Df Sum of Sq     RSS    AIC  Pr(>Chi)
<none>               3029092 3836.2
age      1     51647 3080739 3841.4  0.007213 **
gender   1      7695 3036788 3835.3  0.297932
dialect 22   4328218 7357311 4171.1 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Regression results are shown in Table 1. There was a significant effect of age ($p = .007$) and dialect ($p < .001$); gender was non-significant ($p = .298$).

4. 
```
> TukeyHSD(aov(r), which = c("dialect"))
  Tukey multiple comparisons of means
    95% family-wise confidence level
```

| | Coef. | S.E. | $p(\mathrm{X}^2)$ |
|---|---|---|---|
| (Intercept) | 76.88 | 7.87 | |
| Age | 0.77 | 0.30 | .007 |
| Gender | | | .298 |
| Female | 4.49 | 4.44 | |
| Male | −4.49 | | |
| Dialect | | | $< .001$ |
| ... | | | |

Table 1: Results of the regression model; dialect coefficients are omitted for reasons of space.

```
Fit: aov(formula = r)

$dialect
                                                  diff  ...
Boston-Atlantic Provinces                 -3.574486e+01  ...
...
```

There are significant differences (at $\alpha_{fw} = .05$) between many pairs of dialects.

5. The New York City dialect has the largest distance between (o) and (oh) in the data, according to the model; at the mean age, this dialect has a mean distance of roughly 580, 401 over the overall sample mean. And, according to the Tukey post-hoc tests, it is significantly different than every other dialect except Providence, another known region of resistance to the merger.

# 2 Multicollinearity

```
> d <- read.table(
         "http://www.wellformedness.com/courses/LING82100/Data/LCV-SES.tsv",
         header = TRUE
)
```

1. This is non-trivial multicollinearity:

```
> n <- ncol(d)
> for (n1 in 1:(n - 1)) {
+     x <- d[, n1]
+     x.name <- names(d)[n1]
+     for (n2 in (n1 + 1):n) {
```

```
+         y <- d[, n2]
+         y.name <- names(d)[n2]
+         r <- cor(x, y)
+         cat(sprintf("r(%s, %s) = %+.3f\n", x.name, y.name, r))
+     }
+ }
r(occ, res) = +0.575
r(occ, sc1) = +0.677
r(occ, sc2) = +0.536
r(res, sc1) = +0.343
r(res, sc2) = +0.252
r(sc1, sc2) = +0.754
```

2. After standardization, the variables should each have $\bar{X} = 0$ and $s = 1$, and they do.

```
> for (name in names(d)) {
+     x <- d[[name]]
+     x <- (x - mean(x)) / sd(x)
+     d[[name]] <- x  # Writes it back into the data frame.
+     cat(sprintf("%s: Xbar = %+.3f, s = %.3f\n", name, mean(x), sd(x)))
+ }
occ: Xbar = +0.000, s = 1.000
res: Xbar = -0.000, s = 1.000
sc1: Xbar = +0.000, s = 1.000
sc2: Xbar = -0.000, s = 1.000
```

3. After residualization,, all pairs of variables should have zero correlation, and they do:

```
> d$res <- residuals(lm(res ~ occ, data = d))
> d$sc1 <- residuals(lm(sc1 ~ occ + res, data = d))
> d$sc2 <- residuals(lm(sc2 ~ occ + res + sc1, data = d))
> n <- ncol(d)
> for (n1 in 1:(n - 1)) {
+     x <- d[, n1]
+     x.name <- names(d)[n1]
+     for (n2 in (n1 + 1):n) {
+         y <- d[, n2]
+         y.name <- names(d)[n2]
+         r <- cor(x, y)
+         cat(sprintf("r(%s, %s) = %+.3f\n", x.name, y.name, r))
+     }
+ }
r(occ, res) = +0.000
r(occ, sc1) = +0.000
```

```
r(occ, sc2) = -0.000
r(res, sc1) = -0.000
r(res, sc2) = +0.000
r(sc1, sc2) = -0.000
```