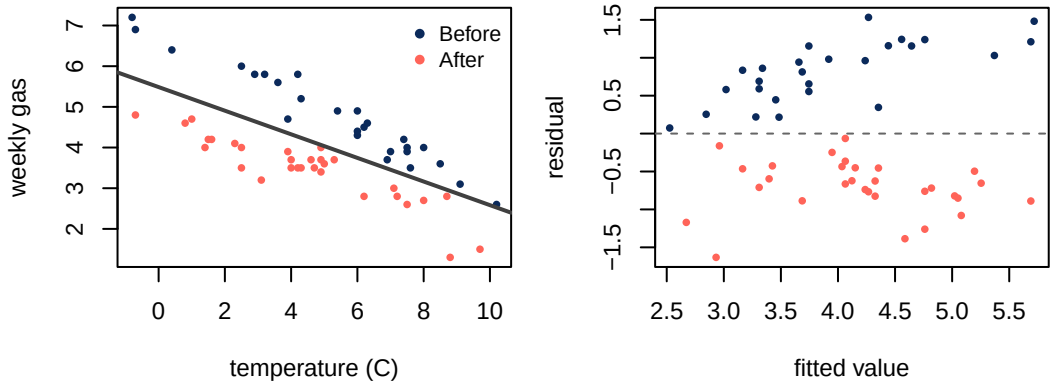
Handout 9

STAT3530 Applied Linear Models

This week we will cover categorical predictors, which require special treatment because their values are non-numeric, *e.g.*, red/blue/green, high/low, small/medium/large, on/off, and the like.

Does insulation reduce energy use?

Weekly gas use and average outside temperature at one house over 56 weeks. Between heating seasons the home underwent cavity-wall insulation, so some observations are recorded *before* insulation, and others *after*. We might want to use this to estimate the effect of insulation on gas use.



1. A simple linear regression of gas on temperature misses obvious structure. Interpret the pattern in the residual plot.

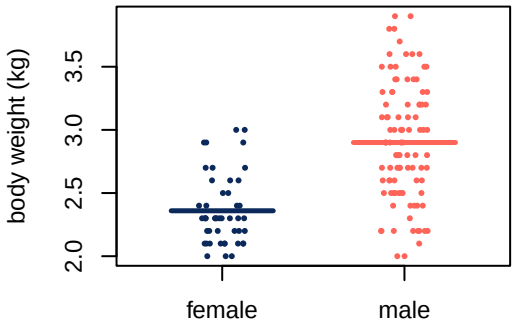
The `insulation` variable in the dataset is categorical, with levels `before` and `after`.

2. Ignoring for the moment that the values of `insulation` are non-numeric, propose a model that includes temperature and insulation as predictors.

$$\text{gas}_i =$$

SLR with a single categorical predictor

Here we'll look at a simple linear regression model with a single categorical predictor having two levels: body weight of 144 cats, 47 female and 97 male.



3. Propose a model for the data, writing `bwt` for body weight and `sex` for sex, ignoring for the moment that the values of `sex` are non-numeric.

$$\text{bwt}_i =$$

4. Now define an *encoding* for the `sex` variable that makes the model well-defined.

$$\text{sex}_i = \begin{cases} \quad , & \text{if cat } i \text{ is F} \\ \\ \quad , & \text{if cat } i \text{ is M} \end{cases}$$

5. Write the model equation for each group and interpret the parameters.

$$\text{female}: \quad \mathbb{E}[\text{bwt}] = \qquad\qquad \text{male}: \quad \mathbb{E}[\text{bwt}] =$$

6. Refer to the midterm if needed. The least squares estimates are:

$$\hat{\beta}_0 = \qquad\qquad \hat{\beta}_1 =$$

7. So inference on β_1 is equivalent to...

Factors in R

Now let's look at how R handles the encoding and model fitting.

Categorical variables are stored as one of two data types in R: they are either *factors* or *character strings*.

```
# string
str(cats$sex.chr)
```

```
 chr [1:144] "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" ...
```

```
# factor
str(cats$sex.fct)
```

```
 Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
```

`lm()` handles both.

```
lm(bwt ~ sex.fct, data = cats) |> coef()
```

```
(Intercept)    sex.fctM
  2.3595745   0.5404255
```

8. Using the output above, give the mean body weight for each group.

female = _____ kg male = _____ kg

R can encode factors using indicator variables. The output below shows a single indicator variable derived from the factor that encodes `F` as the **reference level**.

```
contrasts(cats$sex.fct)
```

```
  M
F 0
M 1
```

9. Why not encode the factor without a reference level and specify instead one indicator per level as shown below? Explain what goes wrong.

$$\text{bwt}_i = \beta_0 + \beta_1 \text{F}_i + \beta_2 \text{M}_i + \epsilon_i \quad \text{so that} \quad X = \begin{bmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix}$$

An energy use model

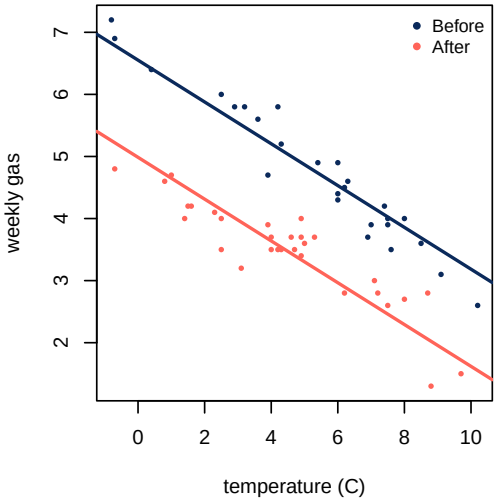
The output below showsthe indicator encoding of the `insul` variable (insulation) and a fitte model.

```
contrasts(whiteside$insul)
```

```
       After
Before     0
After      1
```

```
fit <- lm(gas ~ temp + insul, data = whiteside)
coef(fit)
```

```
(Intercept)        temp   insulAfter
   6.551329   -0.336697    -1.565205
```



10. What is the reference level?

11. Rewrite the proposed model and define an appropriate indicator for the `insul` predictor.

12. This is a *single* model describing two parallel lines, one per factor level. Write the model equation for each factor level, and interpret the model parameters.

    insul $= 0$: insul $= 1$:

13. Interpret the confidence interval below and identify the assumptions needed for a causal interpretation.

```
confint(fit)['insulAfter', ]
```

```
    2.5 %    97.5 %
-1.759869 -1.370540
```

Adding interactions

Here we consider including an **interaction** between `temp` and `insul` by adding a product term to the model.

$$\text{gas}_i = \beta_0 + \beta_1 \text{temp}_i + \beta_2 \text{insul}_i + \beta_3 (\text{temp}_i \times \text{insul}_i) + \epsilon_i$$

The output below shows some example rows of the predictor matrix. Note that the fourth column is exactly a product of the `temp` predictor and the `insul` indicator.

```
model.matrix(gas ~ temp * insul, data = whiteside)[c(1, 2, 27, 28), ]
```

```
   (Intercept) temp insulAfter temp:insulAfter
1            1 -0.8          0             0.0
2            1 -0.7          0             0.0
27           1 -0.7          1            -0.7
28           1  0.8          1             0.8
```

14. Rewrite the model equation by group and interpret the model parameters.

```
$$\text{insul} = 0: \hspace{2in} \qquad \text{insul} = 1: \hspace{2in}$$
```

The output below shows the fitted model.

```
fit.int <- lm(gas ~ temp * insul, data = whiteside)
round(summary(fit.int)$coefficients, 4)
```

```
                Estimate Std. Error  t value Pr(>|t|)
(Intercept)       6.8538     0.1360  50.4091    0e+00
temp             -0.3932     0.0225 -17.4874    0e+00
insulAfter       -2.1300     0.1801 -11.8272    0e+00
temp:insulAfter   0.1153     0.0321   3.5907    7e-04
```

15. We can formally test whether the the slopes differ via the partial test on β_3. State the hypotheses and explain the two possible interpretations.

16. With an interaction in the model, the difference in means depends on temperature. Fix temperature at t and write the difference in means under the model.

$$\mathbb{E}[\text{gas} \mid \text{insul} = 1] - \mathbb{E}[\text{gas} \mid \text{insul} = 0] =$$

17. Rewrite this as a contrast $c^T\beta$.

$$c^T = \begin{bmatrix} ____ & ____ & ____ & ____ \end{bmatrix}$$

The output below shows confidence intervals for the contrast at a few values of t.

```
at <- function(t) {
  fit <- lm(gas ~ I(temp - t) * insul, data = whiteside)
  confint(fit)['insulAfter', ]
}

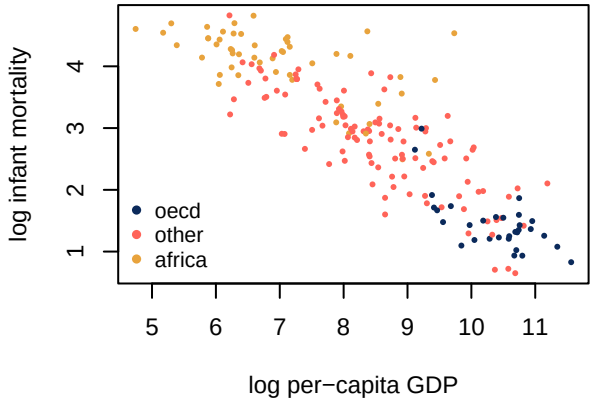
tvals <- c(0, 2.5, 5, 7.5, 10)
round(cbind(t = tvals, t(sapply(tvals, at))), 3)
```

```
        t  2.5 % 97.5 %
[1,]  0.0 -2.491 -1.769
[2,]  2.5 -2.076 -1.608
[3,]  5.0 -1.730 -1.377
[4,]  7.5 -1.508 -1.022
[5,] 10.0 -1.350 -0.604
```

12. Using the intervals, explain the effect of insulation on mean weekly gas use.

More than two levels

Infant mortality and per-capita GDP for 193 countries: 31 OECD, 52 African, 110 other. Both variables are on the log scale. Does infant mortality differ by region after accounting for national income?



Inspect the encoding of `group` for this dataset: now there are two indicator variables.

```
contrasts(un$group)
```

```
       other africa
oecd       0      0
other      1      0
africa     0      1
```

13. Identify the reference level.

14. Write the indicators.

$$\text{other} = \begin{cases} \quad , & \text{if} \\ \quad , & \text{if} \end{cases} \qquad \text{africa} = \begin{cases} \quad , & \text{if} \\ \quad , & \text{if} \end{cases}$$

15. Propose a model without interactions and interpret the parameters (on the log scale).

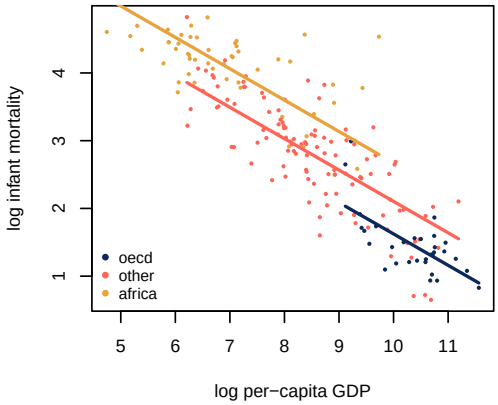
$$\log(\text{mort}_i) =$$

The output below shows the fitted model.

```
lm(lmort ~ lgdp + group, data = un) |> coef()
```

```
(Intercept)        lgdp  groupother groupafrica
  6.2550828  -0.4631826   0.4801703   1.0491120
```

16. Interpret the estimates $e^{0.48} = 1.6$ and $e^{1.05} = 2.9$ in context.



Inference with multiple levels

The partial t tests give pairwise comparisons.

```
fit <- lm(lmort ~ lgdp + group, data = un)
round(summary(fit)$coef, 4)
```

```
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)   6.2551     0.3250  19.2464        0
lgdp         -0.4632     0.0303 -15.2829        0
groupother    0.4802     0.1100   4.3639        0
groupafrica   1.0491     0.1468   7.1478        0
```

17. To check your understanding, interpret the test on the `africa` indicator. Write the hypotheses and report the result of the test in context.

Testing whether the categorical variable matters *at all* requires an F test.

```
# fit a reduced model without the categorical predictor
fit.red <- lm(lmort ~ lgdp, data = un)

# extra sum of squares F test
anova(fit)
```

```
Analysis of Variance Table

Response: lmort
           Df  Sum Sq Mean Sq F value    Pr(>F)
lgdp        1 174.566 174.566 800.034 < 2.2e-16 ***
group       2  12.032   6.016  27.572 3.113e-11 ***
Residuals 189  41.239   0.218
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

17. Using the output above, state the hypotheses and explain why the test uses 2 numerator degrees of freedom.

18. Interpret the result of the test in context.

Adding interactions

19. Allowing the slopes to differ adds one product term per indicator. Fill in the blanks.

$$\text{lmort}_i = \beta_0 + \beta_1 \text{lgdp}_i + \beta_2 \text{oth}_i + \beta_3 \text{afr}_i + \beta_4(_____) + \beta_5(_____) + \epsilon_i$$

20. Complete the table.

group	intercept	slope
`oecd`	β_0	β_1
`other`	_____	_____
`africa`	_____	_____

21. Interpret the parameters β_4 and β_5.

The slopes differ if *either* $\beta_4 \neq 0$ or $\beta_5 \neq 0$, so this is again an F test.

```
fit.int <- lm(lmort ~ lgdp * group, data = un)

anova(fit, fit.int)
```

```
Analysis of Variance Table

Model 1: lmort ~ lgdp + group
Model 2: lmort ~ lgdp * group
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    189 41.239
2    187 38.042  2    3.1972 7.8579 0.0005287 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

22. State the hypotheses of the test above, explain the numerator degrees of freedom, and interpret the result in context.

23. The output below shows the partial t tests. Why are none of the interactions significant even though the F test was?

```
round(summary(fit.int)$coefficients, 4)
```

```
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        6.4764     1.3384  4.8390   0.0000
lgdp              -0.4845     0.1290 -3.7576   0.0002
groupother         0.9429     1.3746  0.6859   0.4936
groupafrica       -0.4005     1.3908 -0.2879   0.7737
lgdp:groupother   -0.0589     0.1340 -0.4392   0.6610
lgdp:groupafrica   0.1966     0.1395  1.4091   0.1605
```

24. The output below shows the model refit with a different reference level. Verify your answer above by identifying the partial t test for the relevant indicator, and interpret the result in context.

```
un2  <- transform(un, group = relevel(group, ref = 'africa'))
fit2 <- lm(lmort ~ lgdp * group, data = un2)

round(summary(fit2)$coef, 4)
```

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.0759     0.3784 16.0573   0.0000
lgdp             -0.2880     0.0532 -5.4094   0.0000
groupoecd         0.4005     1.3908  0.2879   0.7737
groupother        1.3433     0.4913  2.7342   0.0069
lgdp:groupoecd   -0.1966     0.1395 -1.4091   0.1605
lgdp:groupother  -0.2554     0.0645 -3.9607   0.0001
```