ST2304 Continuation Exam Solution

Introduction

This is one solution: it is not the only one, and may not even be the best.

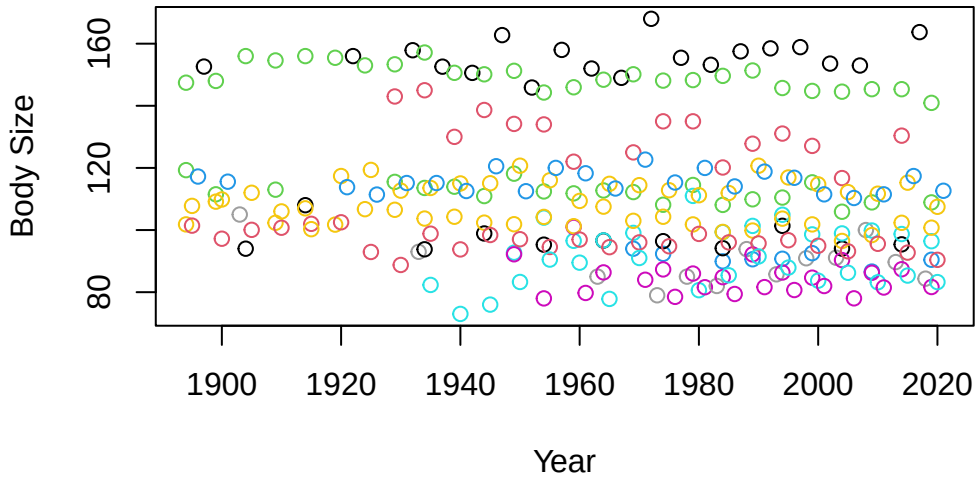
The Problem

One question in the 2023 exam was about changes in body size of birds over time. Here we will look at a different part of the same problem: does body size change in mammals, in particular in the genus *Sorex* (shrews).

Figure 1: An Ornate shrew *Attribution: Pacific Southwest Region U.S. Fish and Wildlife Service, Public domain, via Wikimedia Commons* https://en.wikipedia.org/wiki/Ornate_shrew#/media/File:Sorex_ornatus_relictus.jpg

```
SorexData <- read.csv("~/Dropbox/Teaching/ST2304/ST2304 - 2023/SorexData.csv")
SorexData$Colour <- as.numeric(as.factor(SorexData$Species))

plot(SorexData$Year, SorexData$mean, col=SorexData$Colour,
     xlab="Year", ylab="Body Size")
```



Is there a change over time?

First, from the plot above it looks like some species are larger than others. So we need to include Species as a factor:

```
lmSp <- lm(mean~ Species, data=SorexData)
lmYr <- lm(mean~ Year, data=SorexData)

lm1 <- lm(mean~ Year + Species, data=SorexData)
summary(lm1)
```

```
Call:
lm(formula = mean ~ Year + Species, data = SorexData)

Residuals:
     Min       1Q   Median       3Q      Max
-14.1633  -2.4438   0.0057   2.4787  12.8907
```

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               220.329024  16.560206  13.305  < 2e-16 ***
Year                       -0.032856   0.008408  -3.907  0.00012 ***
SpeciesSorex_cinereus     -59.686933   1.350425 -44.199  < 2e-16 ***
SpeciesSorex_fumeus       -43.723483   1.415251 -30.895  < 2e-16 ***
SpeciesSorex_haydeni      -63.968090   1.896394 -33.731  < 2e-16 ***
SpeciesSorex_hoyi         -71.194642   1.495196 -47.616  < 2e-16 ***
SpeciesSorex_longirostris -74.303101   1.755818 -42.318  < 2e-16 ***
SpeciesSorex_monticolus   -42.960981   1.350425 -31.813  < 2e-16 ***
SpeciesSorex_nanus        -65.648687   1.613583 -40.685  < 2e-16 ***
SpeciesSorex_ornatus      -58.769612   1.693231 -34.709  < 2e-16 ***
SpeciesSorex_pacificus    -24.676497   1.516575 -16.271  < 2e-16 ***
SpeciesSorex_palustris     -6.591204   1.350880  -4.879 1.89e-06 ***
SpeciesSorex_trowbridgii  -40.275368   1.385335 -29.073  < 2e-16 ***
SpeciesSorex_tundrensis   -55.254421   1.582021 -34.926  < 2e-16 ***
SpeciesSorex_ugyunak      -67.641328   1.582327 -42.748  < 2e-16 ***
SpeciesSorex_vagrans      -53.084769   1.361316 -38.995  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.469 on 253 degrees of freedom
Multiple R-squared:  0.9635,    Adjusted R-squared:  0.9613
F-statistic: 444.8 on 15 and 253 DF,  p-value: < 2.2e-16
```

Indeed, species are different sizes: in the model with only Year, the R^2 is 4.2%, and with only species the the R^2 is 96.1%. This leads us to the incredible conclusion that species are different sizes. In practice there is no need to do a test for this: the difference is so huge.

In the model with both Year and species, the Year effect is negative: each year the size decreases by 0.033 mm per year. The 95% confidence interval is a decrease of 0.016 mm - 0.049mm per year.

So it appears that body size does decrease over time.

Although it is not the main focus of the analysis, we can also look at the Species effects. The `(Intercept)` is the intercept for the reference species, which here is *Sorex bendirii*, the Marsh Shrew. The estimate is 220.3mm, i.e. if we believe the model, a Marsh Shrew on the year 0 would be 22cm long. This is either unrealistic or a miracle (but as this species is also called the Jesus shrew…).

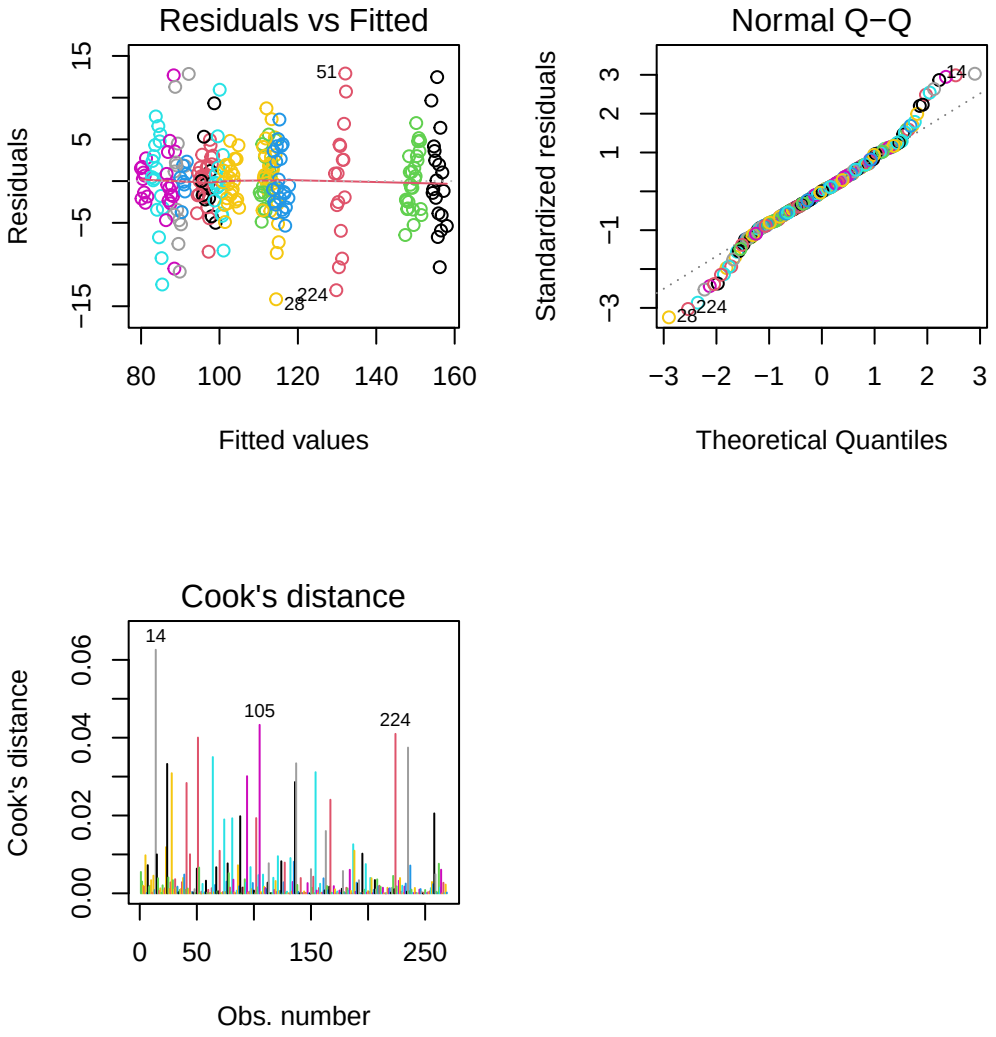
The other estimates are the difference in size from the Marsh Shrew, (for example) *S. vagrans*, the vagrant shrew (or wandering shrew) is 53.1mm smaller than the marsh shrew. Because the Year effect is assume to be the same for every species, this difference is the same for any year:

the model assume that every species gets smaller by 0.033 mm each year, so the difference is always the same.

How good is this model?

First, the R^2 for the model is 96.3%, so the model explains most of the variation in the data. This is, of course, because the species differ so much. We can also check the model fit by looking at the residuals:

```
par(mfrow=c(2,2))
plot(lm1, c(1,2,4), col=SorexData$Colour)
```



Overall these look OK, except the normal probability plot (top right) suggests that the tails of the distribution of the residuals are a bit wide: more like a t distribution than a normal. In a bit more detail:

- the residual plot (top left) looks OK. The only real pattern is the clumping into different groups. This is simply because each group is one (or more) species. This is clear from giving a different colour to each species.
- as already noted, the normal probability plot suggests thick tails. We can't see any outliers or anything else weird.
- the Cook's D values (bottom left) look OK: they are all small.

For the moment the model looks OK, if not perfect.

Does the change vary between species?

We can ask if the change in size varies between species by adding a `Species:Year` interaction. This is the code for the model:

```
lm2 <- lm(mean~ Species*Year, data=SorexData)
# Can also do this with
# lm2 <- update(lm1, .~.+Year:Species)
# summary(lm2)
```

Before looking at the model in detail, we can ask whether the model is worth it. This is a confirmatory analysis, so we can use `anova()` to compare the models with and without the interaction:

```
anova(lm1, lm2)
```

```
Analysis of Variance Table

Model 1: mean ~ Year + Species
Model 2: mean ~ Species * Year
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    253 5051.8
2    239 4278.7 14     773.1 3.0846 0.0002001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This suggests that the data are unlikely if there was no difference between species. The R^2 is now 96.9%, so still high, and has increased by 0.6%. Not a lot, but most of the variation is between species anyway.

What are the estimates?

```
summary(lm2)
```

```
Call:
lm(formula = mean ~ Species * Year, data = SorexData)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5672  -2.2737  -0.0075   2.2081  12.6459

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    69.53278   62.01309   1.121  0.26330
SpeciesSorex_cinereus         122.98088   75.64738   1.626  0.10533
SpeciesSorex_fumeus           151.87024   80.80182   1.880  0.06139 .
SpeciesSorex_haydeni          183.41541  198.56266   0.924  0.35657
SpeciesSorex_hoyi            -108.92807   98.13516  -1.110  0.26812
SpeciesSorex_longirostris      26.89447  185.67876   0.145  0.88496
SpeciesSorex_monticolus       -11.73656   75.64738  -0.155  0.87684
SpeciesSorex_nanus            184.60618   96.45408   1.914  0.05682 .
SpeciesSorex_ornatus           91.67041   95.73007   0.958  0.33924
SpeciesSorex_pacificus        425.55349  101.99333   4.172 4.22e-05 ***
SpeciesSorex_palustris        232.63224   75.63471   3.076  0.00234 **
SpeciesSorex_trowbridgii       66.39398   78.99871   0.840  0.40150
SpeciesSorex_tundrensis         3.28820  118.64967   0.028  0.97791
SpeciesSorex_ugyunak          134.76508  119.51663   1.128  0.26063
SpeciesSorex_vagrans          122.84855   76.71017   1.601  0.11060
Year                            0.04386    0.03154   1.390  0.16570
SpeciesSorex_cinereus:Year     -0.09300    0.03853  -2.414  0.01655 *
SpeciesSorex_fumeus:Year       -0.09955    0.04114  -2.420  0.01626 *
SpeciesSorex_haydeni:Year      -0.12520    0.09980  -1.254  0.21089
SpeciesSorex_hoyi:Year          0.01862    0.04974   0.374  0.70853
SpeciesSorex_longirostris:Year -0.05177    0.09351  -0.554  0.58040
SpeciesSorex_monticolus:Year   -0.01563    0.03853  -0.406  0.68536
SpeciesSorex_nanus:Year        -0.12692    0.04884  -2.599  0.00994 **
SpeciesSorex_ornatus:Year      -0.07653    0.04874  -1.570  0.11772
```

```
SpeciesSorex_pacificus:Year      -0.22874    0.05182  -4.414 1.53e-05 ***
SpeciesSorex_palustris:Year      -0.12191    0.03853  -3.164  0.00176 **
SpeciesSorex_trowbridgii:Year    -0.05425    0.04019  -1.350  0.17832
SpeciesSorex_tundrensis:Year     -0.03023    0.05994  -0.504  0.61444
SpeciesSorex_ugyunak:Year        -0.10271    0.06036  -1.702  0.09013 .
SpeciesSorex_vagrans:Year        -0.08955    0.03907  -2.292  0.02277 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.231 on 239 degrees of freedom
Multiple R-squared:  0.9691,    Adjusted R-squared:  0.9653
F-statistic: 258.1 on 29 and 239 DF,  p-value: < 2.2e-16
```

There are a lot of them! We can see that the `Year` effect is 0.044, i.e. is now positive. This means that *S. bendirii*, our Marsh shrew, might be increasing in size (although the confidence interval for this is quite wide, so we can't be sure if it is increasing or decreasing).

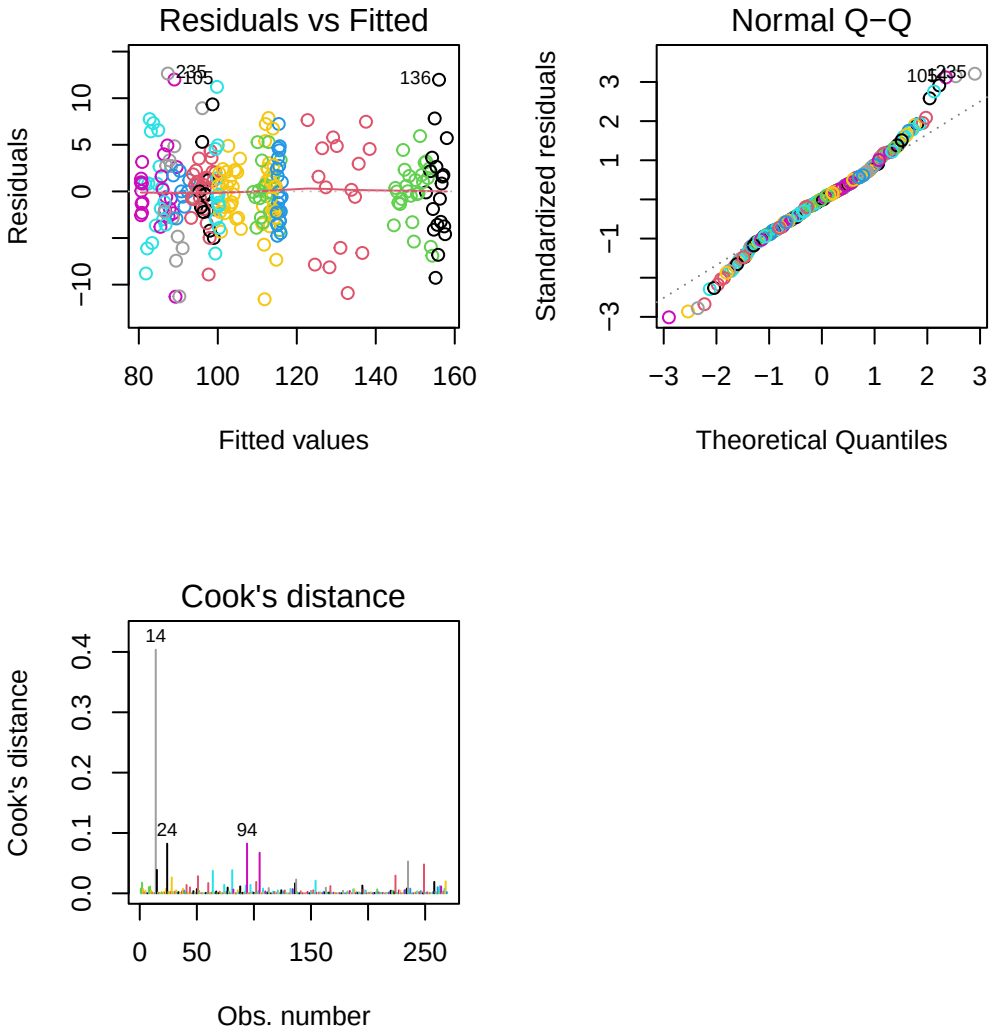
If we look at the other species, we see some of the effects are different. This means they are different from *S. bendirii*. e.g. *S. pacificus* (the Pacific shrew) has changed by 0.044 + -0.229 = -0.185mm per year, i.e. has been getting smaller by about 2mm every decade.

Note that the effect for *S. tundrensis* is negativem because the slope is smaller than it is for *S. bendirii*, but the effect is 0.044 + -0.03 = 0.014mm per year, so it is still positive (although, again the confidence intervals will really say we don't know wht direction the effect is)

how well does the model fit the data?

We already know that the R^2 is high, so it seems to fit well. But let's check the fit of the model with the interactions.

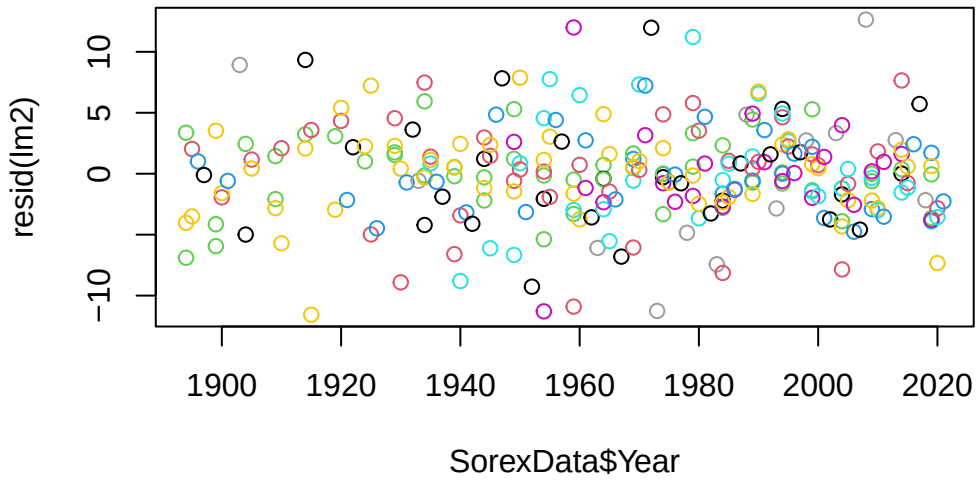
```
par(mfrow=c(2,2))
plot(lm2, c(1,2,4), col=SorexData$Colour)
```



These are very similar to the earlier plots, so our conclusions are the same.

One thing we didn't do earlier is to look at the pattern of residuals over time. We can do this here:

```
plot(SorexData$Year, resid(lm2), col=SorexData$Colour)
```



These look OK: htey seem fairly random.

Conclusions

- there is an effect of time on size in shrews.
- the effect does vary between species, so some might be increasing in size.
- most of the variation is between species
- the model fits well, but the residuals have rather this tails.

We have not covered how to deal with thick tailed models. For this problem I might leave it as it is. The alternative is an approach called **robust regression**, which reduces the influence of the outlying points.

One of the interesting things about this is that most of the variation is between species, which is not what we are interested in. This makes assessment of `Year` difficult, because R^2 is not useful.

One way of assessing the change is to look at the percentage change in a century:

```
Predicts <- expand.grid(Species = unique(SorexData$Species), Year=c(1900,2000))

Predicts$pred <- predict(lm2, newdata = Predicts)
PercentChange <- 100*(Predicts$pred[Predicts$Year==2000]-Predicts$pred[Predicts$Year==1900
names(PercentChange) <- Predicts$Species[Predicts$Year==2000]
round(PercentChange, 1)
```

```
    Sorex_fumeus    Sorex_palustris      Sorex_vagrans     Sorex_cinereus
            -4.8              -5.1               -4.3              -5.0
Sorex_monticolus  Sorex_trowbridgii     Sorex_bendirii        Sorex_nanus
             2.5              -0.9                2.9              -8.6
   Sorex_ornatus    Sorex_pacificus        Sorex_hoyi   Sorex_tundrensis
            -3.3             -12.9                7.9               1.4
   Sorex_ugyunak Sorex_longirostris      Sorex_haydeni
            -6.4              -1.0               -8.3
```

So, for example, our Marsh shrew has got larger by about 3% (if we believe the model!), but the pacific shrew has decreased by 13%! Of course, we shouldput confidence intervals around this too…