Advanced Lab 4: More Linear Modeling and ANOVA

Introduction

We will now start considering more complex models, with more than one predictor. We will start with a model including all the predictors:

```
fitAll <- lm(time~., data=targetingFinal)
summary(fitAll)
```

Let's take a closer look at this model. We have:

- the intercept
- a term for the subject ID. We should probably remove this.
- iat
- a factor level for male gender. Presumably the female case is the base case.
- factor levels for age, with 18 being the base case.
- a factor level for race being white. Presumably the base case is race being black.
- a factor level for the weapon value being set to unarmed, with base case being armed.
- a factor level for action incorrect, with base case being the correct action.

Calling anova on the model will tell us about each factor as a whole, and whether its contribution is statistically significant:

```
anova(fitAll)
```

We see that the contribution with ther largest P-value, and hence smaller effect, is from the subject ID. This is good, there is no reason whatsoever that the subject ID should have anything to do with the data. Let's remove it and reconsider:

```
fitNoSubj <- lm(time~.-subject, data=targetingFinal)
summary(fitNoSubj)
anova(fitNoSubj)
```

We can see that the residual standard error did not change almost at all with the removal of the subject variable. We see a number of other variables that are potentially not influential.

One powerful utility that R offers is the ability to remove one variable at a time from a model, and consider all the resulting smaller models. The drop1 method is one way to proceed for that:

```
drop1(fitNoSubj)
```

What we see is the effect of removing each of the variables, in terms of how much the RSS will change. The AIC column portrays the "Akaike Information Criterion". The AIC is a number that takes into consideration the RSS but also penalizes models based on the number of parameters used. It is technically given by the formula:

$$-2\text{log-likelihood} + 2p$$

A larger model will fit the data better, and so will have a larger log-likelihood and therefore a smaller value for -2log-likelihood. But it will also have more parameters, so $2p$ will be larger. The formula aims to balanced the two: The smaller the AIC the more you have gained by the bigger model, while also accounting for how bigger the model was. A model with smaller AIC is fitting the data "more

economically". There is often an indeterminate additive constant involved in AIC computations, so they are only appropriate for comparisons between models.

The first line, marked <none>, contains the full model, while the remaining lines indicate the effect with the corresponding predictor removed.

In this case we see that the biggest gain in AIC can be obtained by dropping the age variable. This would make sense since that variable, coded as a factor variable, uses 5 parameters. Let's remove it and consider the model again:

```
fitNoSubjNoAge <- fitNoSubj %>% update(formula=.~.-age)
summary(fitNoSubjNoAge)
anova(fitNoSubjNoAge)
drop1(fitNoSubjNoAge)
```

We now see race as a next possible variable to remove.

```
fitNoSubjNoAgeNoRace <- fitNoSubjNoAge %>% update(formula=.~.-race)
summary(fitNoSubjNoAgeNoRace)
anova(fitNoSubjNoAgeNoRace)
drop1(fitNoSubjNoAgeNoRace)
```

Removing gender seems to be the next step:

```
fitNoSubjNoAgeNoRaceNoGender <- fitNoSubjNoAgeNoRace %>% update(formula=.~.-gender)
summary(fitNoSubjNoAgeNoRaceNoGender)
anova(fitNoSubjNoAgeNoRaceNoGender)
drop1(fitNoSubjNoAgeNoRaceNoGender)
```

We now have a model where removing a variable does not provide an AIC improvement.

We can also consider adding models in, one at a time:

```
add1(fitNoSubjNoAgeNoRaceNoGender, scope=~.+race+gender+age)
```

We see that none of them is an improvement.

CLEANUP

We will leave it as is for now as the data did not show any signs of extreme skewness. Here is a starting plot that shows the density distribution for time for armed and unarmed weapons, and with different graphs for each race and gender combination:

```
ggplot(targetingFinal) +
    aes(x=time, color=weapon) +
    geom_density() +
    facet_grid(race~gender)
```

We can see that mean reaction times were slower for the unarmed weapons.

Let us compute some numerical summaries:

```
targetingCorrect %>%
    group_by(race, weapon, gender) %>%
    summarize(mean=mean(time),
                    se=sd(time)/sqrt(n()))
```

We can also plot these:

```
ggplot(targetingCorrect) +
    aes(x=weapon, y=time, color=race) +
    stat_summary(fun.data=mean_se, position=position_dodge(0.2)) +
    facet_wrap(~gender)
```

We probably expected the marked difference in reaction times between armed and unarmed subjects. For female subjects, the race of the subject seems to play a small factor.

```
fit1 <- lm(time~race*weapon, data=targetingCorrect)
summary(fit1)
anova(fit1)
```

We can see a significant overall effect, but we can also see that the interaction terms are not significant. We remove them from the model:

```
fit2 <- lm(time~race+weapon, data=targetingCorrect)
summary(fit2)
anova(fit2)
```

We can compare the two models to see if there are differences, and there is no significant difference:

```
anova(fit1, fit2)
```

We can get some default diagnostics from plotting the fit:

```
par(mfrow=c(2,2))
plot(fit2)
```

The residuals appear to be normal and with constant variance. We can visualize their effect against the other predictors:

```
ggplot(targetingCorrect) +
    aes(x=weapon, y=resid(fit2), color=race) +
ggplot(targetingCorrect) +
    aes(x=race, y=resid(fit2), color=weapon) +
    geom_point(position=position_jitter(0.1))
ggplot(targetingCorrect) +
    aes(x=race, y=resid(fit2), color=weapon) +
    geom_point(position=position_dodge(0.1))
ggplot(targetingCorrect) +
    aes(x=gender, y=resid(fit2), color=interaction(race, weapon)) +
    geom_point(position=position_dodge(0.2))
```

We can look at how iat might be related to those residuals, there's clearly a relation there:

```
ggplot(targetingCorrect) +
    aes(x=iat, y=resid(fit2), color=weapon) +
    geom_point() +
    geom_smooth()
```

Now let's add the subject's gender into the model:

```
fit3 <- lm(time~race+weapon+gender+race:gender+weapon:gender, data=targetingCorrect)
summary(fit3)
anova(fit3)
anova(fit2, fit3)
```

We see that the subject's gender does not appear to be significant.

Finally, we look at whether we should remove race from the model as well:

```
fit4 <- lm(time~weapon, data=targetingCorrect)
summary(fit4)
anova(fit4)
anova(fit4, fit3)
```

```
fit5 <- lm(time~poly(iat, 2)+weapon, data=targetingCorrect)
summary(fit5)
anova(fit5)
```

ggplot(targetingCorrect) + aes(x=race, y=time, color=gender) + geom_point() + geom_line(aes(group=subject))