Regression with categorical variables

Packages for this section

```
library(tidyverse)
library(broom)
```

The pigs revisited

- Recall pig feed data, after we tidied it:

```
my_url <- "http://www.utsc.utoronto.ca/~butler/c32/pigs2.txt"
pigs <- read_delim(my_url, " ")
pigs
```

pig	feed	weight
1	feed1	60.8
2	feed1	57.0
3	feed1	65.0
4	feed1	58.6
5	feed1	61.7
1	feed2	68.7
2	feed2	67.7
3	feed2	74.0
4	feed2	66.3
5	feed2	60.8

Summaries

```
pigs %>%
  group_by(feed) %>%
  summarize(n = n(), mean_wt = mean(weight),
            sd_wt = sd(weight))
```

feed	n	mean_wt	sd_wt
feed1	5	60.62	3.064637
feed2	5	69.30	2.926602
feed3	5	94.10	3.613170
feed4	5	86.24	2.896204

Running through aov and lm

- What happens if we run this through lm rather than aov?
- Recall aov first:

```
pigs.1 <- aov(weight ~ feed, data = pigs)
summary(pigs.1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## feed         3   3521  1173.5   119.1 3.72e-11 ***
## Residuals   16    158     9.8
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

and now `lm`

```
pigs.2 <- lm(weight ~ feed, data = pigs)
tidy(pigs.2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	60.62	1.403567	43.189962	0.0000000
feedfeed2	8.68	1.984943	4.372921	0.0004731
feedfeed3	33.48	1.984943	16.866980	0.0000000
feedfeed4	25.62	1.984943	12.907170	0.0000000

```
glance(pigs.2)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC
0.9571521	0.9491181	3.138471	119.1379	0	4	-49.02205	108.0441

Understanding those slopes

- Get one slope for each category of categorical variable feed, except for first.
- feed1 treated as "baseline", others measured relative to that.
- Thus prediction for feed 1 is intercept, 60.62 (mean weight for feed 1).
- Prediction for feed 2 is $60.62 + 8.68 = 69.30$ (mean weight for feed 2).
- Or, mean weight for feed 2 is 8.68 bigger than for feed 1.
- Mean weight for feed 3 is 33.48 bigger than for feed 1.
- Slopes can be negative, if mean for a feed had been smaller than for feed 1.

Reproducing the ANOVA

- Pass the fitted model object into `anova`:

```
anova(pigs.2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
feed	3	3520.526	1173.508	119.1379	0
Residuals	16	157.600	9.850	NA	NA

- Same as before.
- But no Tukey this way:

```
TukeyHSD(pigs.2)
```

```
## Error in UseMethod("TukeyHSD"): no applicable method for 'TukeyHSD' appl
```

The crickets

- Male crickets rub their wings together to produce a chirping sound.
- Rate of chirping, called "pulse rate", depends on species and possibly on temperature.
- Sample of crickets of two species' pulse rates measured; temperature also recorded.
- Does pulse rate differ for species, especially when temperature accounted for?

The crickets data

Read the data:

```
my_url="http://www.utsc.utoronto.ca/~butler/c32/crickets2.csv"
crickets <- read_csv(my_url)
crickets %>% sample_n(10)
```

species	temperature	pulse_rate
exclamationis	20.8	67.9
exclamationis	24.0	80.4
niveus	22.1	60.7
niveus	18.3	49.6
niveus	18.3	47.6
niveus	26.5	77.0
exclamationis	24.0	78.7
niveus	18.9	50.3
niveus	26.5	77.7
niveus	23.5	69.8

Fit model with `lm`

```
crickets.1 <- lm(pulse_rate ~ temperature + species,
                 data = crickets)
```

Can I remove anything? No:

```
drop1(crickets.1, test = "F") %>%  select(-Df, -`Sum of Sq`)
```

	RSS	AIC	F value	Pr(>F)
	89.34987	38.81575	NA	NA
temperature	4465.43244	158.07416	1371.3541	0
species	687.35382	100.06472	187.3994	0

`drop1` is right thing to use in a regression with categorical (explanatory) variables in it: "can I remove this categorical variable *as a whole*?"

The summary

```
glance(crickets.1)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik
0.9895889	0.9888453	1.786356	1330.719	0	3	-60.39497

```
tidy(crickets.1)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-7.210906	2.5509378	-2.826767	0.0085824
temperature	3.602753	0.0972881	37.031799	0.0000000
speciesniveus	-10.065291	0.7352622	-13.689390	0.0000000

Conclusions

- Slope for temperature says that increasing temperature by 1 degree increases pulse rate by 3.6 (same for both species)
- Slope for `speciesniveus` says that pulse rate for niveus about 10 lower than that for exclamationis at same temperature (latter species is baseline).
- R-squared of almost 0.99 is very high, so that the prediction of pulse rate from species and temperature is very good.

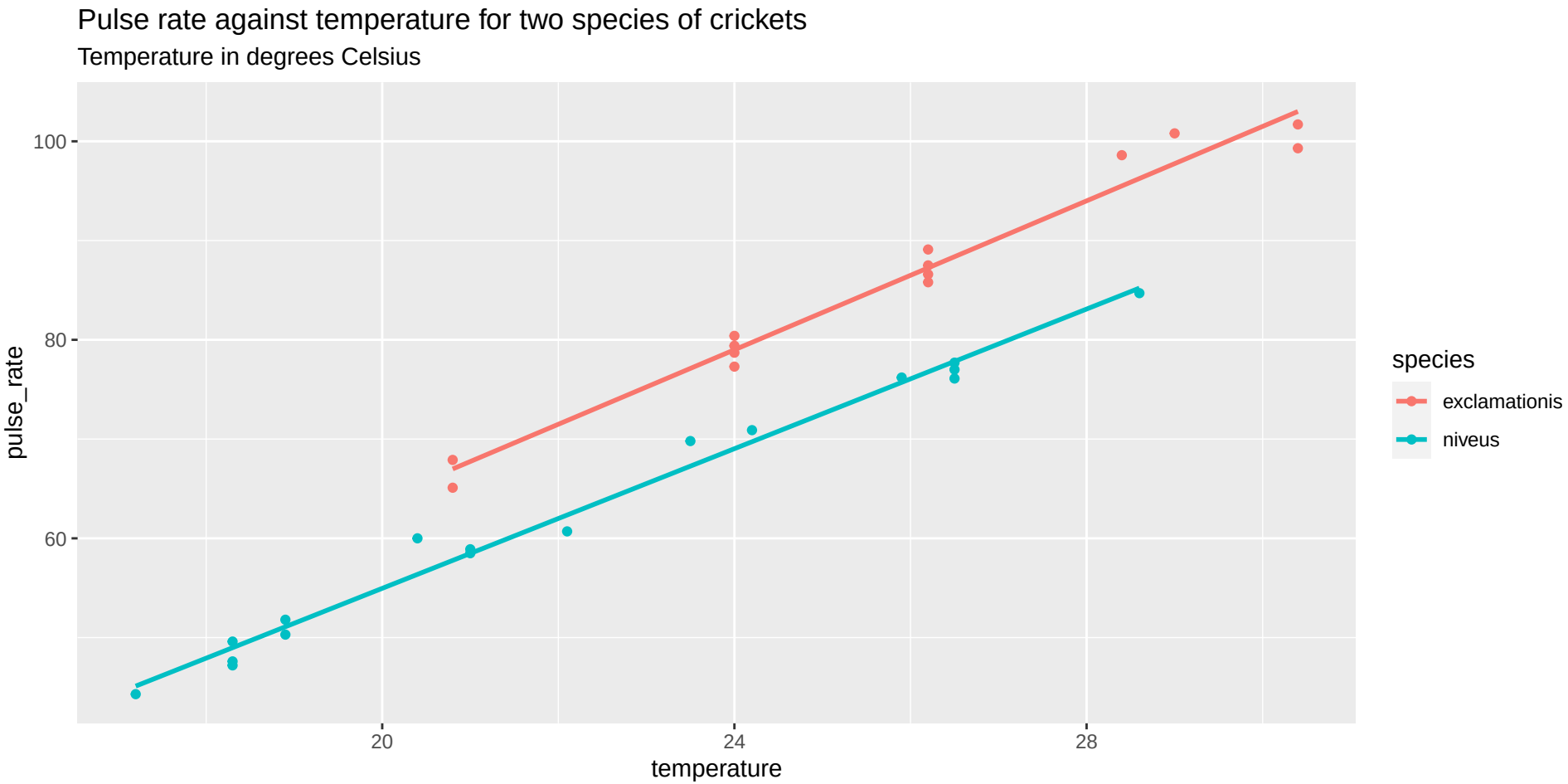
To end with a graph

- Two quantitative variables and one categorical: scatterplot with categories distinguished by colour.
- This graph seems to need a title, which I define first.

```
t1 <- "Pulse rate against temperature for two species of crickets"
t2 <- "Temperature in degrees Celsius"
```

```
ggplot(crickets, aes(x = temperature, y = pulse_rate,
  colour = species)) +
  geom_point() + geom_smooth(method = "lm", se = F) +
  ggtitle(t1, t2)
```

The graph


Pulse rate against temperature for two species of crickets
Temperature in degrees Celsius
pulse_rate
100
80
60
20
24
28
temperature
species
exclamationis
niveus