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://ritsokiguess.site/datafiles/pigs2.txt"
pigs <- read_delim(my_url, " ")
pigs</pre>
```

# A	tibb]	Le: 20	х З
	pig	feed	weight
	<dbl></dbl>	<chr></chr>	<dbl></dbl>
1	1	feed1	60.8
2	2	feed1	57
3	3	feed1	65
4	4	feed1	58.6
5	5	feed1	61.7
6	1	feed2	68.7
7	2	feed2	67.7
8	3	feed2	74
9	4	feed2	66.3
10	5	feed2	69.8
11	1	feed3	92.6

Summaries

#	A tibl	ole: 4	x 4	
	feed	n	mean_wt	sd_wt
	< chr >	<int></int>	<dbl></dbl>	<dbl></dbl>
1	feed1	5	60.6	3.06
2	feed2	5	69.3	2.93
3	feed3	5	94.1	3.61
4	feed4	5	86.2	2.90

Running through aov and lm

• What happens if we run this through 1m rather than aov?

Recall aov first:

pigs.1 <- aov(weight ~ feed, data = pigs)
summary(pigs.1)</pre>

Df Sum Sq Mean Sq F value Pr(>F) feed 3 3521 1173.5 119.1 3.72e-11 *** Residuals 16 158 9.9 ---Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' :

and now lm

```
pigs.2 <- lm(weight ~ feed, data = pigs)
summary(pigs.2)</pre>
```

Call: lm(formula = weight ~ feed, data = pigs) Residuals. Min 10 Median 30 Max -3.900 -2.025 -0.570 1.845 5.000 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 60.620 1.404 43.190 < 2e-16 *** feedfeed2 8.680 1.985 4.373 0.000473 *** feedfeed3 33.480 1.985 16.867 1.30e-11 *** feedfeed4 25.620 1.985 12.907 7.11e-10 *** ___ Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.138 on 16 degrees of freedom

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)
```

```
Analysis of Variance Table
```

```
Response: weight

Df Sum Sq Mean Sq F value Pr(>F)

feed 3 3520.5 1173.51 119.14 3.72e-11 ***

Residuals 16 157.6 9.85

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

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

Error in UseMethod("TukeyHSD"): no applicable method for 'TukeyHSD' applied to an o

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://ritsokiguess.site/datafiles/crickets2.csv"
crickets <- read_csv(my_url)
crickets %>% slice_sample(n = 10) # display sample of rows
```

```
# A tibble: 10 x 3
```

	species	temperature	pulse_rate
	<chr></chr>	<dbl></dbl>	<dbl></dbl>
1	niveus	18.3	47.6
2	exclamationis	26.2	89.1
3	exclamationis	30.4	102.
4	exclamationis	24	79.4
5	exclamationis	30.4	99.3
6	niveus	20.4	60
7	niveus	18.3	49.6
8	niveus	26.5	77.7
9	niveus	21	58.5
10	exclamationis	26.2	86.6

Fit model with 1m

Can I remove anything? No:

```
drop1(crickets.1, test = "F")
```

```
Single term deletions
```

Model: pulse_rate ~ temperature + species Df Sum of Sq RSS AIC F value Pr(>F) <none> 89.3 38.816 temperature 1 4376.1 4465.4 158.074 1371.4 < 2.2e-16 *** species 1 598.0 687.4 100.065 187.4 6.272e-14 *** ---Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' :

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

summary(crickets.1)

```
Call:
lm(formula = pulse_rate ~ temperature + species, data = crickets)
Residuals:
   Min 10 Median 30
                                 Max
-3.0128 -1.1296 -0.3912 0.9650 3.7800
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.21091 2.55094 -2.827 0.00858 **
temperature 3.60275 0.09729 37.032 < 2e-16 ***
speciesniveus -10.06529 0.73526 -13.689 6.27e-14 ***
___
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.786 on 28 degrees of freedom
Multiple R-squared: 0.9896, Adjusted R-squared: 0.9888
F-statistic: 1331 on 2 and 28 DF, p-value: < 2.2e-16
```

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 crick
t2 <- "Temperature in degrees Celsius"
ggplot(crickets, aes(x = temperature, y = pulse_rate,
    colour = species)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE) +
    ggtitle(t1, subtitle = t2) -> g
```

The graph



