

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
```

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

Summaries

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

```
# A tibble: 4 x 4  
  feed      n mean_wt sd_wt  
  <chr> <int>   <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 `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	9.9		

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

and now lm

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

Call:

```
lm(formula = weight ~ feed, data = pigs)
```

Residuals:

Min	1Q	Median	3Q	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 <chr>	temperature <dbl>	pulse_rate <dbl>
1	niveus	21	58.9
2	exclamationis	28.4	98.6
3	niveus	17.2	44.3
4	exclamationis	20.8	67.9
5	niveus	18.3	47.6
6	niveus	26.5	76.1
7	niveus	28.6	84.7
8	niveus	18.9	50.3
9	niveus	25.9	76.2
10	niveus	20.4	60

Fit model with `lm`

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

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 ' ' 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	1Q	Median	3Q	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.

What this model is doing: a graph

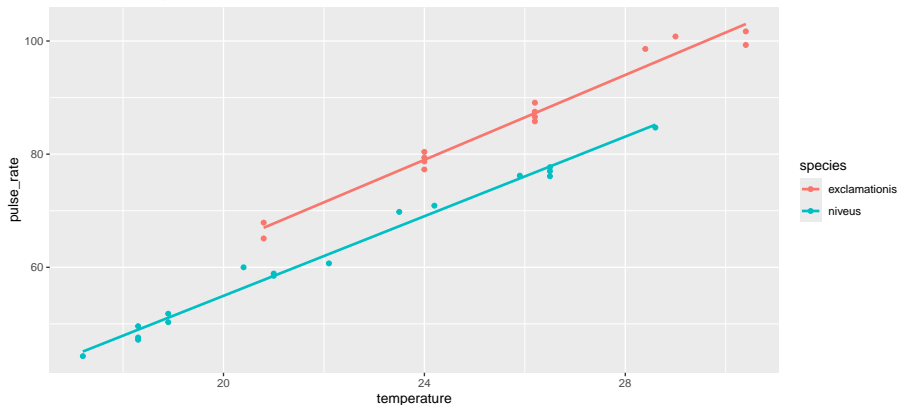
- Two quantitative variables and one categorical: scatterplot with categories distinguished by colour.
- This graph seems to need a title, which I define first.

```
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

09

Pulse rate against temperature for two species of crickets
Temperature in degrees Celsius



Residuals

- We can also check residuals from this kind of model:

```
crickets.1a <- augment(crickets.1)
crickets.1a
```

```
# A tibble: 31 x 9
```

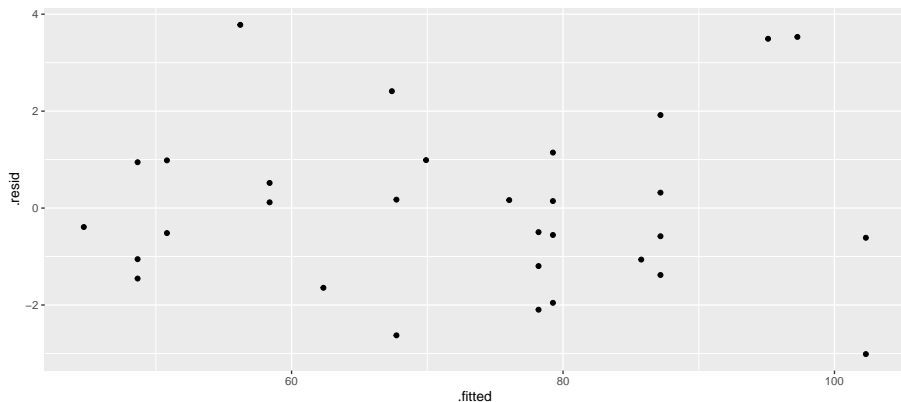
	pulse_rate	temperature	species	.fitted	.resid	.hat	.sigma
	<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	67.9	20.8	exclamationis	67.7	0.174	0.144	1.83
2	65.1	20.8	exclamationis	67.7	-2.63	0.144	1.74
3	77.3	24	exclamationis	79.3	-1.96	0.0806	1.78
4	78.7	24	exclamationis	79.3	-0.555	0.0806	1.83
5	79.4	24	exclamationis	79.3	0.145	0.0806	1.83
6	80.4	24	exclamationis	79.3	1.14	0.0806	1.80
7	85.8	26.2	exclamationis	87.2	-1.38	0.0720	1.80
8	86.6	26.2	exclamationis	87.2	-0.581	0.0720	1.83
9	87.5	26.2	exclamationis	87.2	0.319	0.0720	1.83
10	89.1	26.2	exclamationis	87.2	1.92	0.0720	1.78

```
# i 21 more rows
```

```
# i 1 more variable: .std.resid <dbl>
```

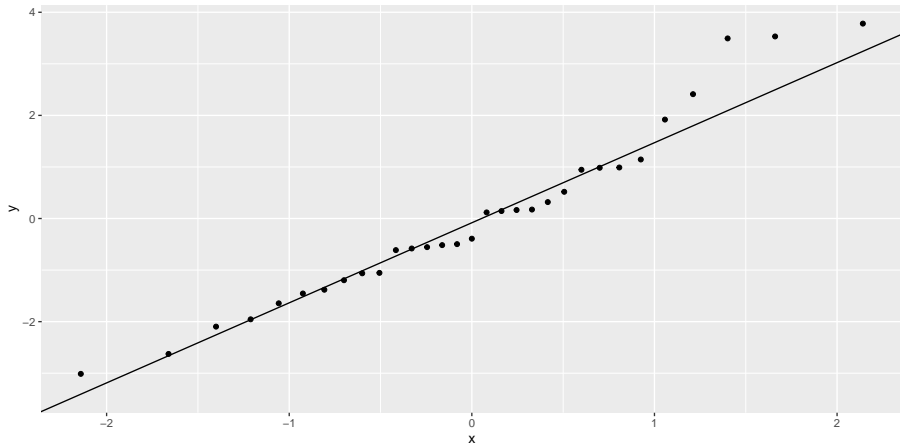
Residuals vs fitted

```
ggplot(crickets.1a, aes(x = .fitted, y = .resid)) +  
  geom_point()
```



Normal quantile plot of residuals

```
ggplot(cricket1a, aes(sample = .resid)) +  
  stat_qq() + stat_qq_line()
```



Residuals vs. x 's

doesn't work as you expect:

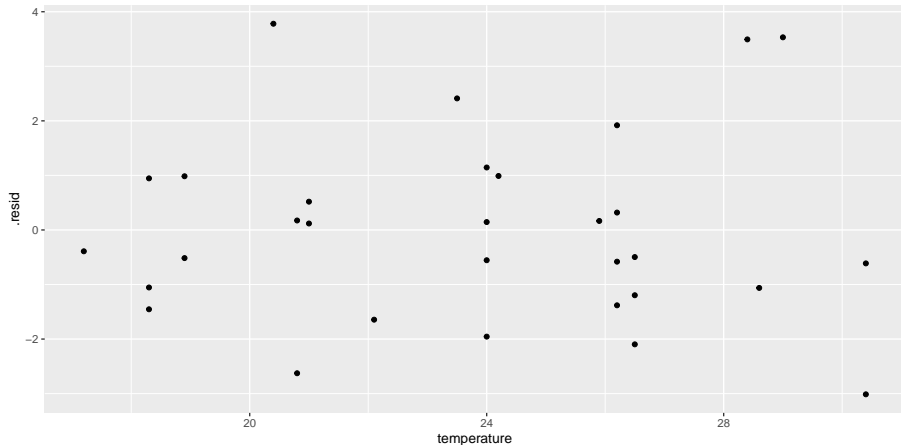
```
crickets.1a %>%  
  pivot_longer(temperature:species,  
               names_to = "xname", values_to = "xval")
```

Error in `pivot_longer()`:

! Can't combine `temperature` <double> and `species` <character>

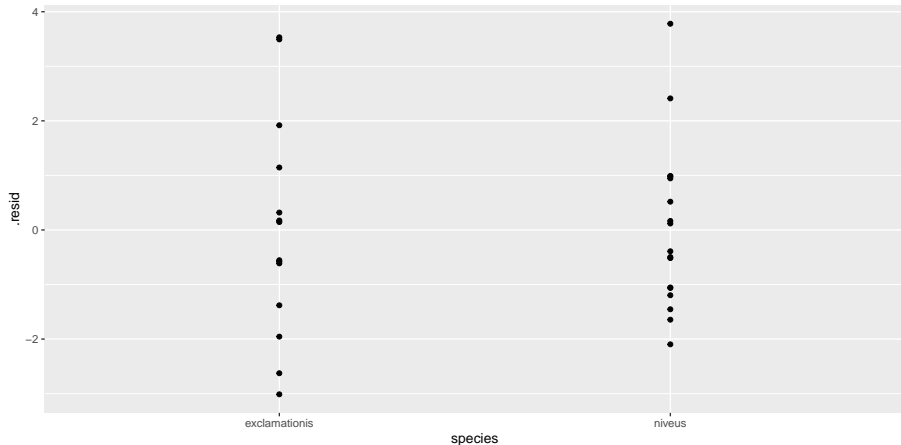
so do them separately 1/3

```
ggplot(crickets.1a, aes(x = temperature, y = .resid)) +  
  geom_point()
```



so do them separately 2/3

```
ggplot(crickets.1a, aes(x = species, y = .resid)) +  
  geom_point()
```



so do them separately 3/3

or perhaps better for categorical x 's

```
ggplot(crickets.1a, aes(x = species, y = .resid)) +  
  geom_boxplot()
```

