Lab 03: Coffee grades

Inference for simple linear regression using mathematical models

Dav King

2022-09-26

Setup

Load packages and data:

```
library(tidyverse)
library(tidymodels)
library(knitr)
```

```
coffee <- read_csv("data/coffee-grades.csv")
```

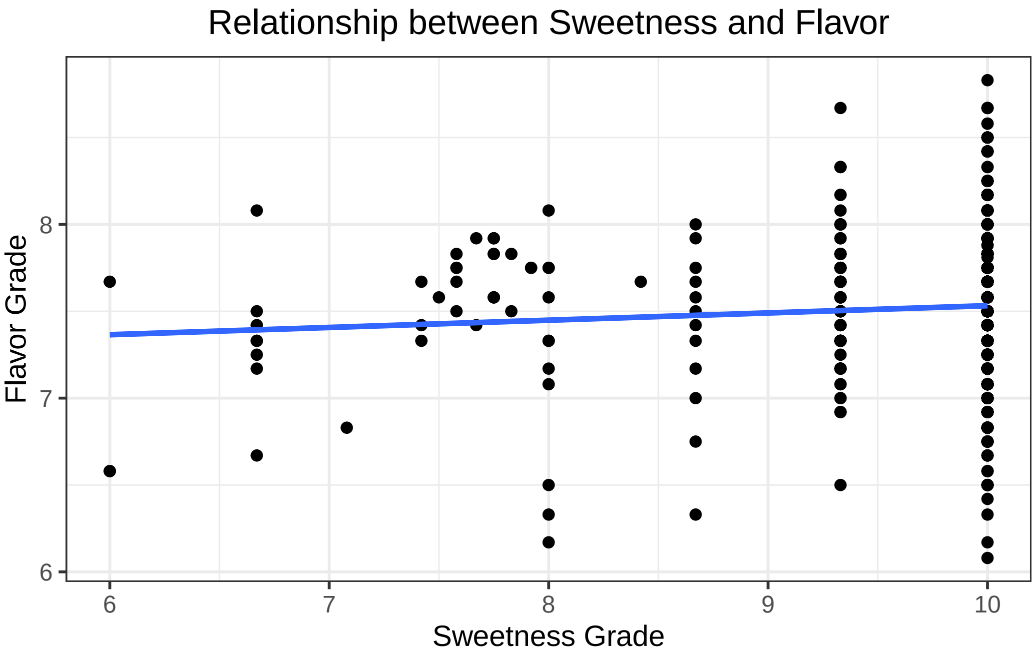
[Select this page for the "Workflow & formatting" and "Reproducible report" sections in Gradescope.]

Exercises

Exercise 1

```
ggplot(coffee, aes(x = sweetness, y = flavor)) +
  geom_point() +
  geom_smooth(method = "lm", se = F) +
  theme_bw() +
  labs(x = "Sweetness Grade", y = "Flavor Grade",
       title = "Relationship between Sweetness and Flavor") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
`geom_smooth()` using formula 'y ~ x'
```



Because these data are discrete, we see strange column clumping of the data. However, there is a slight positive linear relationship between the two variables (though I'd be hesitant to call that a relationship at all, and it is likely statistically insignificant). The flavor grade seems to range only from 6 to 9, and there is increasing variability in flavor grade as sweetness grade increases.

Exercise 2

```r
sweetModel <- linear_reg() %>%
  set_engine("lm") %>%
  fit(flavor ~ sweetness, data = coffee)

tidy(sweetModel, conf.in = T, conf.level = 0.98) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	7.114	0.183	38.890	0.000	6.688	7.540
sweetness	0.042	0.019	2.257	0.024	-0.001	0.085

…

Exercise 3

For every 1 point increase in Sweetness grade, we would expect to see an increase of 0.042 in the coffee's Flavor grade, on average.

If I were a coffee drinker, I would probably be fine drinking a cup of coffee represented by the intercept - while it corresponds to a sweetness grade of 0, many people will tell you that coffee is not supposed to be sweet anyway, and it still has a flavor score of 7.114 (which is not much lower than a cup of coffee with a sweetness grade of 10, anyway).

...

Exercise 4

```
glance(sweetModel$fit)$sigma
```

```
[1] 0.3401849
```

The average distance that a coffee's flavor score falls from the linear regression line predicting the relationship between sweetness and flavor grades is .3402 - in other words, our predictions are wrong by .3402 flavor grades on average.

…

Exercise 5

$H_0 : \beta_1 = 0$, or the slope of the relationship between sweetness and flavor grades is equal to 0. $H_a : \beta_1 \neq 0$, or the slope of the relationship between sweetness and flavor grades is not equal to 0.

Our test statistic is 2.257, which says that the chance of obtaining a slope of the relationship between sweetness and flavor grades as extreme as ours is 2.257 standard errors away from the mean value of the normal distribution of such slopes under H_0.

The p-value was calculated using a t-distribution of slopes centered around the mean slope under H_0, which here happens to be 0, with n-2 = 1335 degrees of freedom.

At a very conveniently chosen α level of .02, we fail to reject H_0. Our p-value of .024 does not provide sufficient evidence to say that the true population slope of the relationship between sweetness and flavor grades is significantly different from 0.

...

Exercise 6

```r
qt(0.98, df = nrow(coffee) - 2)
```

```
[1] 2.055758
```

Our critical value used to calculate the confidence interval was 2.056.

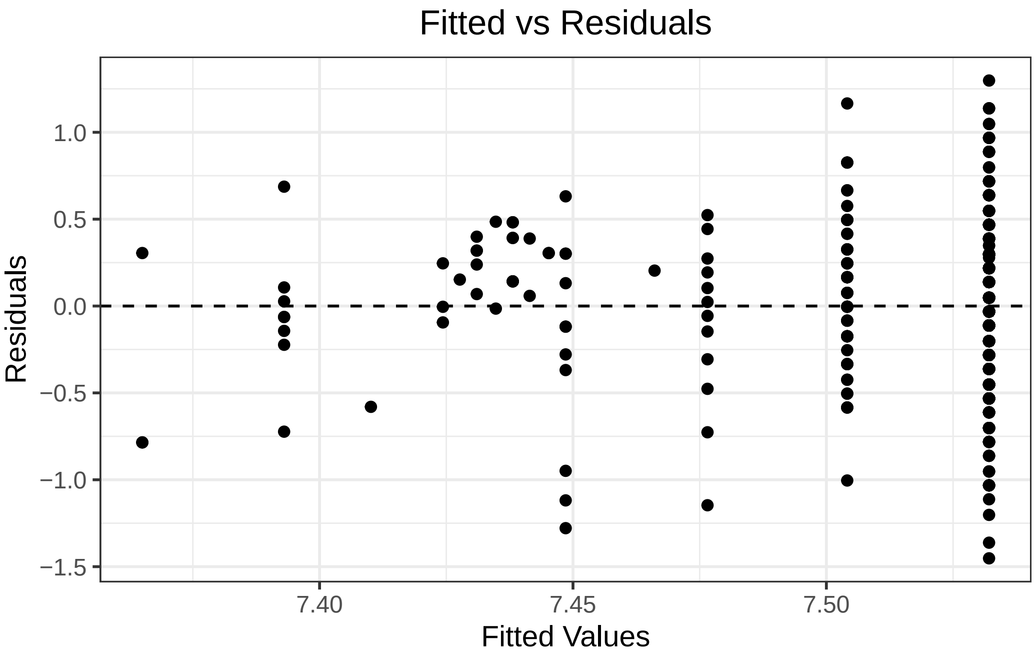
Yes, the confidence interval is consistent with the conclusions from the hypothesis test because the confidence interval contains 0, our null hypothesis. If a confidence interval contains H_0, we fail to reject it at that confidence level.

…

Exercise 7

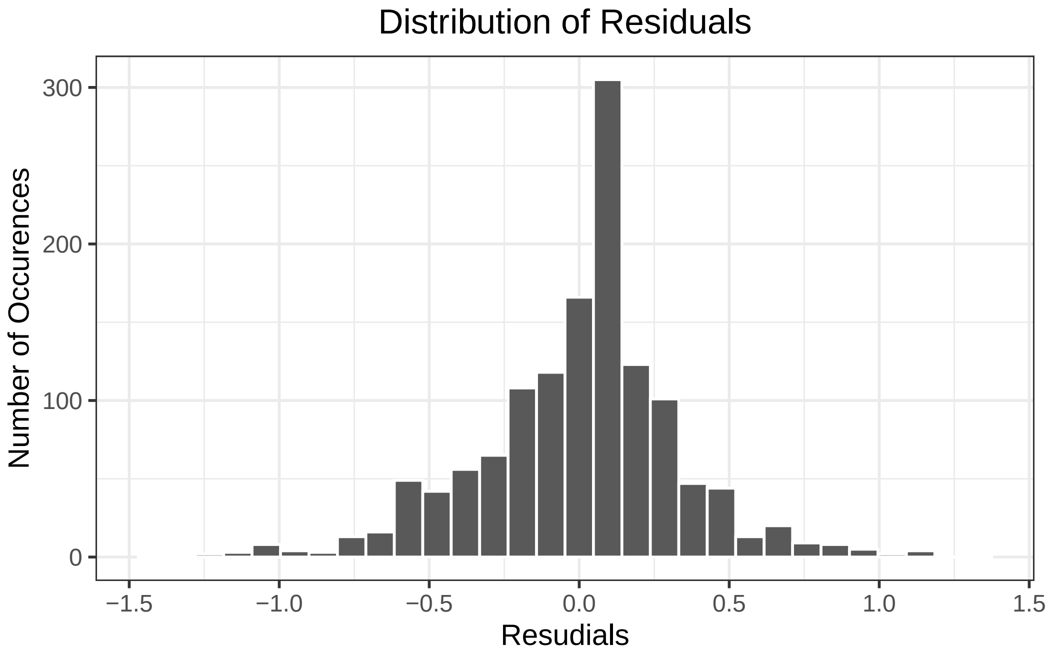
```
augModel <- augment(sweetModel$fit)

ggplot(augModel, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Fitted Values", y = "Residuals", title = "Fitted vs Residuals") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```



```
ggplot(augModel, aes(x = .resid)) +
  geom_histogram(color = "white") +
  theme_bw() +
  labs(x = "Resudials", y = "Number of Occurences",
       title = "Distribution of Residuals") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Distribution of Residuals
Number of Occurences
300
200
100
0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Resudials

...

Exercise 8

The linearity condition is satisfied. When viewing the plot of residuals versus fitted values, there is no clear non-linear trend in the data - the residuals are more or less randomly scattered.

The constant variance condition is not satisfied. At higher fitted values, the residuals begin to scatter more, and there is no real horizontal boundary at which we can contain all of our residuals (because the line would be diagonal instead).

...

Exercise 9

The normality condition is satisfied - we have 1337 different observations in the data, which is well above our threshold of N = 30.

...

Exercise 10

Because we are told in the lab documentation that the coffees can be reasonably treated as a random sample, we can safely say that the independence condition is met.

...