Lecture 13: Regression Inference

Taylor Arnold

regression inference

We have seen how to use the `lm_basic` function to fit models for the mean of some response. We have used both a single mean for the entire dataset as well as multiple means based on a second categorical variable. What happens if we use the same set-up but instead use a numeric variable to predict the value of some response? The output is surprisingly similar, but the interpretation of the results differ slightly.

regression inference

As an example, let's predict the amount each mammal is awake as a function of how many hours it has of rem sleep:

```
model <- lm_basic(awake ~ 1 + sleep_rem, data = msleep)
reg_table(model, level = 0.95)
```

regression inference

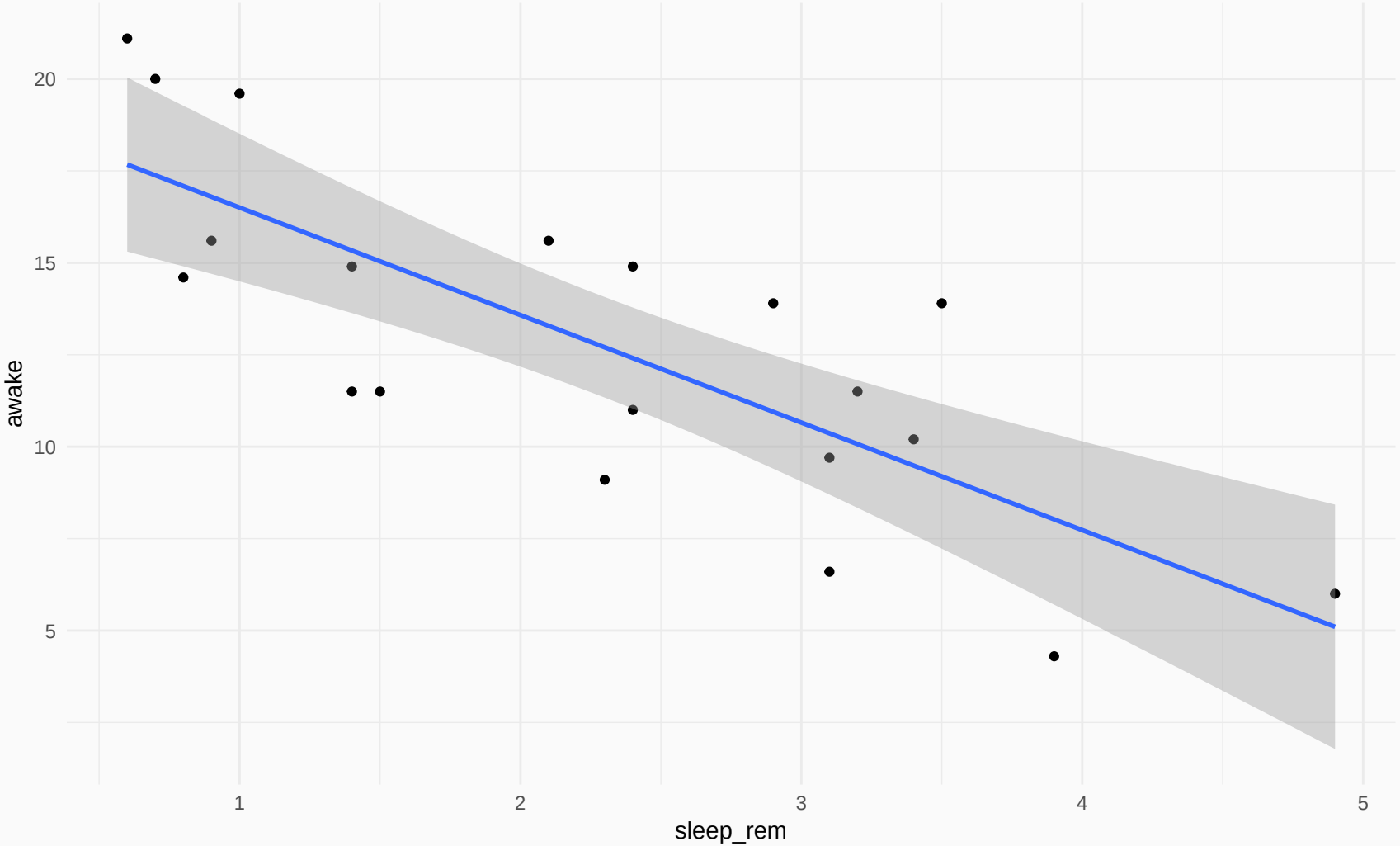
```
model <- lm_basic(awake ~ 1 + sleep_rem, data = msleep)
reg_table(model, level = 0.95)
```

```
##
## Call:
## lm_basic(formula = awake ~ 1 + sleep_rem, data = msleep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8330 -2.7505  0.1404  2.5230  4.7062
##
## Coefficients:
##             Estimate  2.5 % 97.5 %
## (Intercept)   19.426 16.466  22.39
## sleep_rem     -2.923 -4.077  -1.77
##
## Residual standard error: 2.91 on 18 degrees of freedom
## Multiple R-squared:  0.6115, Adjusted R-squared:  0.5899
```

There is once again an intercept term and a row of the table corresponding to the new variable `sleep_rem`. What do these numbers mean? It turns out that this is simply describing a best-fit line through the data. We have already seen how to do this graphically with `geom_smooth`. The line here is, exactly, the line given in this plot:

```
qplot(sleep_rem, awake, data = msleep) + geom_smooth(metho
```

regression inference



regression inference

The `reg_table` function is just giving us the intercept and slope of this line, along with confidence interval bounds for both.

multiple variables

Further, and finally, we can add multiple variables into a single regression. It is even possible to mix continuous and categorical variables into the same model:

```
model <- lm_basic(awake ~ 1 + sleep_rem + vore, data = msle
reg_table(model, level = 0.95)
```

multiple variables

```
model <- lm_basic(awake ~ 1 + sleep_rem + vore, data = msleep)
reg_table(model, level = 0.95)
```

```
##
## Call:
## lm_basic(formula = awake ~ 1 + sleep_rem + vore, data = msleep)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.554 -2.459  0.285  2.716  4.034
##
## Coefficients:
##             Estimate   2.5 % 97.5 %
## (Intercept)  20.3459 14.2320 26.460
## sleep_rem    -3.1563 -4.7079 -1.605
## voreherbi    -0.8790 -5.8266  4.069
## voreinsecti  -0.9271 -6.9977  5.144
## voreomni      0.5673 -4.2954  5.430
##
```

multiple variables

The interpretation becomes, in this case, the change we would expect to see in the response given a *marginal* change in one of the explanatory variables on the right-hand side of the model. That is, how do we expect the mean to change if we modify one (and only one) of the other variables.