AE 19: Linear Regression II

Dav King

3/22/2022

Coming up

- Lab Due Friday at 11:59 PM.

Main ideas

- Review and expand upon concepts from our first regression class.
- Discuss inference for linear regression.

Packages

Please recall

- Response Variable: Variable whose behavior or variation you are trying to understand, on the y-axis. Also called the dependent variable.
- Explanatory Variable: Other variables that you want to use to explain the variation in the response, on the x-axis. Also called independent variables, predictors, or features.
- Predicted value: Output of the model function
  - The model function gives the typical value of the response variable conditioning on the explanatory variables (what does this mean?)
- Residuals: Shows how far each case is from its predicted value
  - Residual = Observed value - Predicted value
  - Tells how far above/below the model function each case is

The linear model with a single predictor

- We're interested in the β_0 (population parameter for the intercept) and the β_1 (population parameter for the slope) in the following model:

$$\hat{y} = \beta_0 + \beta_1 \ x + \epsilon$$

- Unfortunately, we can't get these values
- So we use sample statistics to estimate them:

$$\hat{y} = b_0 + b_1 \ x$$

Least squares regression

The regression line minimizes the sum of squared residuals.

- **Residuals**: $e_i = y_i - \hat{y}_i$,
- The regression line minimizes $\sum_{i=1}^{n} e_i^2$.
- Equivalently, minimizing $\sum_{i=1}^{n}[y_i - (b_0 + b_1\ x_i)]^2$

Returning to our candy model

Let's modify the sugar variable again for easier interpretation.

```
candy_rankings <- candy_rankings %>%
  mutate(sugarpercent100 = sugarpercent*100)
```

Let's run the model again. What units are the explanatory and response variable in? The sugar variable is in percentiles and the win percentage is a perecentage.

```
sugarwins <- linear_reg() %>%
  set_engine("lm") %>%
  fit(winpercent ~ sugarpercent100, data = candy_rankings)
tidy(sugarwins)
```

```
## # A tibble: 2 x 5
##   term            estimate std.error statistic  p.value
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       44.6      3.09       14.5  2.20e-24
## 2 sugarpercent100    0.119    0.0556      2.14 3.49e- 2
```

Interpretation:

Why does having the sugar variable on a scale from 0 to 100 lead to a better interpretation than if we had the scale go from 0 to 1?

Another way to interpret linear regression coefficients is by looking at a *one standard deviation* increase rather than a one unit increase.

Question: Why might you want to look at this instead of a one unit increase?

This allows you to look at a realistic increase that may be more appropriate for the actual scale of the variable. You won't always do this, but it is one option that is sometimes warranted.

You can find a one standard deviation increase by multiplying the coefficient for that term by the standard deviation.

Recall this model:

$$\widehat{WinPercent} = 44.6 + 0.12\ SugarPercentile$$

What is the standard deviation of the sugar percentile variable?

```
candy_rankings %>%
  summarize(sd(sugarpercent100))
```

```
## # A tibble: 1 x 1
##   `sd(sugarpercent100)`
##                   <dbl>
## 1                  28.3
```

Now, let's multiply it by the coefficient here.

```
28.3 * .12
```

```
## [1] 3.396
```

Interpretation: A one percentile increase in sugar content, on average, translates to a 0.119 percentage point increase in wins.

Question: Would you always want to use this sort of interpretation? Does it make sense to do a one standard deviation increase interpretation when working with a dummy variable?

No- with a dummy variable, a one standard deviation increase would make no sense because the only options for a variable are 0 or 1.

Unit increases

Variables will **not always be measured in percentages**. Other potential units include:

- Dollars spent eating out.
- Points scored by a basketball team.
- Votes received by bills in Congress.

These are not the only potential units. It is important to know what units your variables are measured in when interpreting them.

Assessing the quality of the fit

- The strength of the fit of a linear model is commonly evaluated using R^2.
- It tells us what percentage of the variability in the response variable is explained by the model. The remainder of the variability is unexplained.
- R^2 is sometimes called the coefficient of determination.
- **Question**: What does "explained variability in the response variable" mean?

This tells us how much of the ups and downs in the response variable the model can explain.

Obtaining R^2 in R

```
glance(sugarwins) %>%
  pull(r.squared)
```

```
## [1] 0.05251002
```

Roughly 5% of the variability in the percent of time a candy bar wins can be explained by the sugar percentile.

Is this a high or low R^2 value?

This is low- while evaulating whether an R^2 is low or high depends upon the context, this would be considered low across the board.

Question: What is the correlation between the sugar percentile and winning percentage variables? How does this number relate to the R^2 value here?

```
candy_rankings %>%
  summarize(cor(sugarpercent100, winpercent))
```

```
## # A tibble: 1 x 1
##   ‘cor(sugarpercent100, winpercent)‘
##                                <dbl>
## 1                              0.229
```

In a bivariate linear regression, the R^2 is simply the correlation squraed.

R^2 - first principles

- We can write explained variation using the following ratio of sums of squares:

$$R^2 = 1 - \left(\frac{SS\_Error}{SS\_Total}\right)$$

where SS_{Error} is the sum of squared residuals and SS_{Total} is the total variance in the response variable.

Next class, when we talk about multiple regression, we will discuss another measure of model fit called Adjusted R^2 that is preferable for models with many explanatory variables.

CLT Based Inference for Linear Regression

- Population model:

$$\hat{y} = \beta_0 + \beta_1 \ x_1 + \epsilon$$

- Sample model that we use to estimate the population model:

$$\hat{y} = b_0 + b_1 \ x_1$$

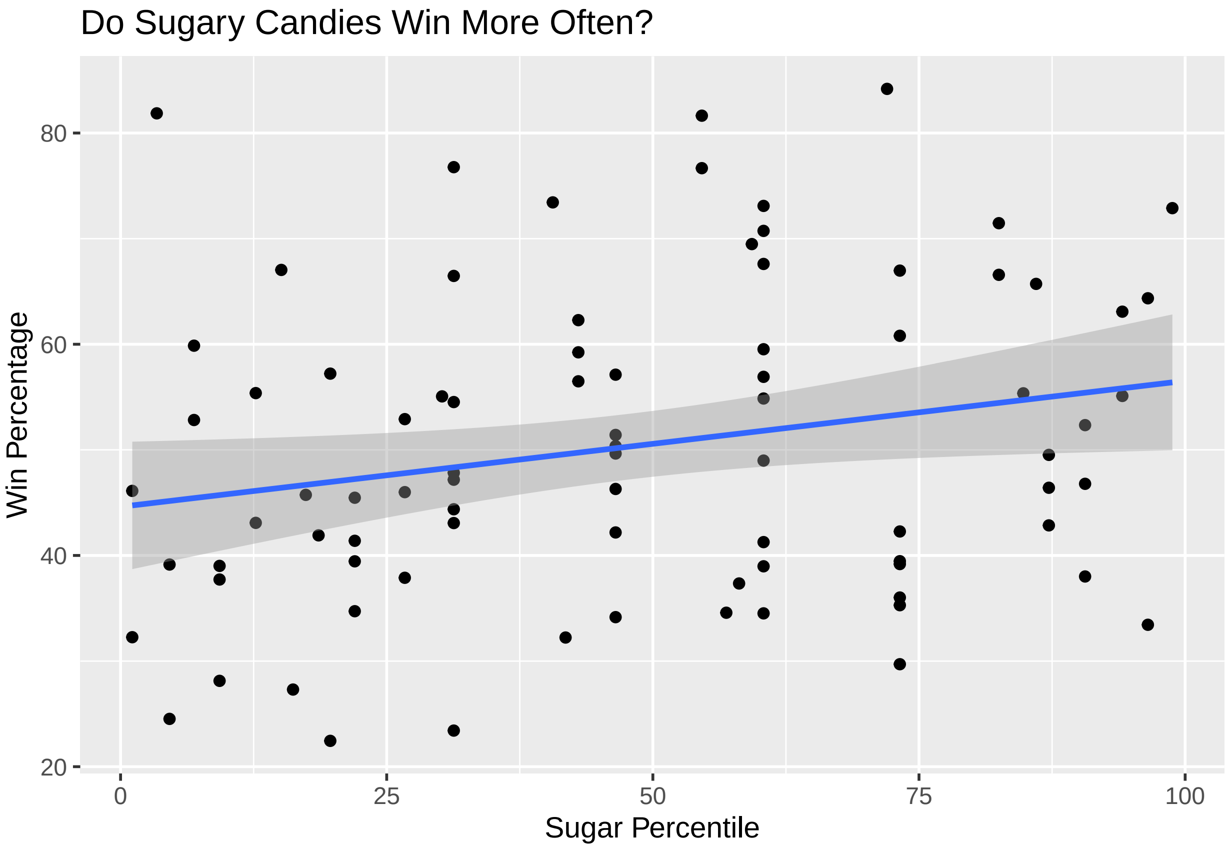
Similar to other sample statistics (mean, proportion, etc) there is variability in our estimates of the slope and intercept.

- Do we have convincing evidence that the true linear model has a non-zero slope?
- What is a confidence interval for the population regression coefficient?

Let's return to our sugar content model:

```
ggplot(data = candy_rankings, aes(x = sugarpercent100, y = winpercent)) +
  labs(title = "Do Sugary Candies Win More Often?", x = "Sugar Percentile", y =
         "Win Percentage") +
  geom_point() +
  geom_smooth(method = "lm")
```

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



Tidy Confidence Interval

$$point\ estimate \pm critical\ value \times SE$$

$$b_1 \pm t^*_{n-2} \times SE_{b_1}$$

```
tidy(sugarwins)
```

```
## # A tibble: 2 x 5
##   term            estimate std.error statistic  p.value
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       44.6      3.09       14.5  2.20e-24
## 2 sugarpercent100    0.119    0.0556      2.14 3.49e- 2
```

Tidy Confidence Interval

A 95% confidence interval for β_1 is given by

```
tidy(sugarwins, conf.int = TRUE, conf.level = 0.95) %>%
  filter(term == "sugarpercent100") %>%
   select(starts_with("conf"))
```

```
## # A tibble: 1 x 2
##   conf.low conf.high
##      <dbl>     <dbl>
## 1  0.00866     0.230
```

non-Tidy Confidence Interval

Using the model output directly and function `qt()` we can also obtain the 95% confidence interval.

```
tidy(sugarwins, conf.int = TRUE, conf.level = 0.95)
```

```
## # A tibble: 2 x 7
##   term            estimate std.error statistic  p.value conf.low conf.high
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)       44.6      3.09       14.5  2.20e-24 38.5        50.7
## 2 sugarpercent100    0.119    0.0556      2.14 3.49e- 2  0.00866     0.230
```

```
0.119 + c(-1, 1) * qt(0.975, 83) * 0.0556
```

```
## [1] 0.008413836 0.229586164
```

But the `tidy()` / `confint_tidy()` methods from `broom` are preferred when constructing confidence intervals.

Interpretation

We are 95% confident that for every additional percentile of sugar, the win percentage is expected to increase, on average, between 0.009 and 0.23 percentage points.

Hypothesis testing for β_1

Is there convincing evidence, based on our sample data, that sugar content is associated with winning percentage?

We can set this up as a hypothesis test, with the hypotheses below.

$H_0 : \beta_1 = 0$. There is no relationship, the slope is 0.

$H_A : \beta_1 \neq 0$. There is a relationship between sugar content and winning percentage.

We only reject H_0 in favor of H_A if the data provide strong evidence that the true slope parameter is different from zero.

Hypothesis testing for β_1

$$T = \frac{b_k - 0}{SE_{b_k}} \sim t_{n-2}$$

The p-values in the `tidy()` output represent the two-sided p-value associated with the observed statistic

$$H_0 : \beta_k = 0 \qquad H_A : \beta_k \neq 0$$

```
tidy(sugarwins)
```

```
## # A tibble: 2 x 5
##   term            estimate std.error statistic  p.value
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       44.6      3.09       14.5  2.20e-24
## 2 sugarpercent100    0.119    0.0556      2.14 3.49e- 2
```

Based on the result above we reject H_0 in favor of H_A. We have enough evidence to suggest that there is an association between sugar content and win percentage.

Practice

For the practice problems, we will use the `endorsements` dataset that FiveThirtyEight compiled. The variables in this dataset are described in more detail here. This dataset includes data on the endorsements that presidential candidates received in the primaries since 1980.

1. Please run a model and interpret a model with the number of endorsements as the explanatory variable and the percentage of votes received as the response variable. Use both the one-unit increase interpretation and one standard deviation increase interpretation.

```
endorsementsmodel <- linear_reg() %>%
  set_engine("lm") %>%
  fit(primary_vote_percentage ~ endorsement_points, data = endorsements)
tidy(endorsementsmodel)
```

```
## # A tibble: 2 x 5
##   term               estimate std.error statistic  p.value
##   <chr>                 <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)           6.60      1.51       4.38 2.76e- 5
## 2 endorsement_points    0.239     0.0306     7.81 4.09e-12
```

```
sd(endorsements$endorsement_points) * 0.239
```

```
## [1] 11.01285
```

For every unit increase in weighted endorsements, we expect to see a 0.239 point increase in primary vote percentage. For every standard deviation increase in weighted endorsements, we expect to see an 11.01285 point increase in primary vote percentage.

Slope: 0.239

Intercept: 6.60

2. Please find and interpret the R^2 for this model.

```
glance(endorsementsmodel) %>%
  pull(r.squared)
```

```
## [1] 0.3632575
```

36.326% of the variance in primary vote percentage can be explained by weighted endorsements.

3. Find and interpret the 95% confidence interval for β_1.

```
tidy(endorsementsmodel, conf.int = T, conf.level = 0.95) %>%
  filter(term == "endorsement_points") %>%
  select(starts_with("conf"))
```

```
## # A tibble: 1 x 2
##   conf.low conf.high
##      <dbl>     <dbl>
## 1    0.179     0.300
```

We are 95% confident that for each unit increase in weighted endorsement points, a candidate would expect to experience an increase in primary vote percentage of between 0.179 and 0.300.