HAHN_Lab2

Hope Hahn

01/23/24

Today we will be continuing the pumpkin case study from last week. We will be using the data that you cleaned and split last time (pumpkins_train) and will be comparing our results today to those you have already obtained. Open and run your Lab 1.Rmd as a first step so those objects are available in your Environment.

Once you have done that, we'll start today's lab by specifying a recipe for a polynomial model. First we specify a recipe that identifies our variables and data, converts the package variable to a numerical form, and then adds a polynomial effect with step_poly()

```
# Specify a recipe
poly_pumpkins_recipe <-
  recipe(price ~ package, data = pumpkins_train) %>%
  step_integer(all_predictors(), zero_based = TRUE) %>%
  step_poly(all_predictors(), degree = 4)
```

How did that work? Later we will learn about model tuning that will let us do things like find the optimal value for degree. For now, we'd like to have a flexible model, so we'll use a relatively large value.

Polynomial regression is still linear regression, so our model specification looks similar to before.

```
# Create a model specification called poly_spec
poly_spec <- linear_reg() %>%
  set_engine("lm") %>%
  set_mode("regression")
```

Question 1: Now take the recipe and model specification that just created and bundle them into a workflow called poly_wf.

```
# Bundle recipe and model spec into a workflow
poly_wf <- workflow() %>%
  add_recipe(poly_pumpkins_recipe) %>%
  add_model(poly_spec)
```

Question 2: fit a model to the pumpkins_train data using your workflow and assign it to poly_wf_fit

```
# Create a model
poly_wf_fit <- poly_wf %>%
  fit(data = pumpkins_train)
```

```
# Print learned model coefficients
poly_wf_fit
```

```
## == Workflow [trained] ==========================================================
## Preprocessor: Recipe
## Model: linear_reg()
##
## -- Preprocessor ----------------------------------------------------------------
## 2 Recipe Steps
##
## * step_integer()
## * step_poly()
##
## -- Model ----------------------------------------------------------------------
##
## Call:
## stats::lm(formula = ..y ~ ., data = data)
##
## Coefficients:
##    (Intercept)  package_poly_1  package_poly_2  package_poly_3  package_poly_4
##        27.9706        103.8566       -110.9068        -62.6442          0.2677
```

```
# Make price predictions on test data
poly_results <- poly_wf_fit %>%
  predict(new_data = pumpkins_test) %>%
  bind_cols(pumpkins_test %>%
  select(c(package, price))) %>%
  relocate(.pred, .after = last_col())

# Print the results
poly_results %>%
  slice_head(n = 10)
```

```
## # A tibble: 10 x 3
##    package price .pred
##      <int> <dbl> <dbl>
##  1       0  13.6  15.9
##  2       0  16.4  15.9
##  3       0  16.4  15.9
##  4       0  13.6  15.9
##  5       0  15.5  15.9
##  6       0  16.4  15.9
##  7       2  34    34.4
##  8       2  30    34.4
##  9       2  30    34.4
## 10       2  34    34.4
```

Now let's evaluate how the model performed on the test_set using yardstick::metrics().

```
# evaluate model performance
metrics(data = poly_results, truth = price, estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       3.27
## 2 rsq     standard       0.892
## 3 mae     standard       2.35
```

Question 3: How do the performance metrics differ between the linear model from last week and the polynomial model we fit today? Which model performs better on predicting the price of different packages of pumpkins?

- **The metrics show that there is less error and a better model fit for the polynomial model. The polynomial model does a better job predicting the price of different packages of pumpkins.**

Let's visualize our model results. First prep the results by binding the encoded package variable to them.

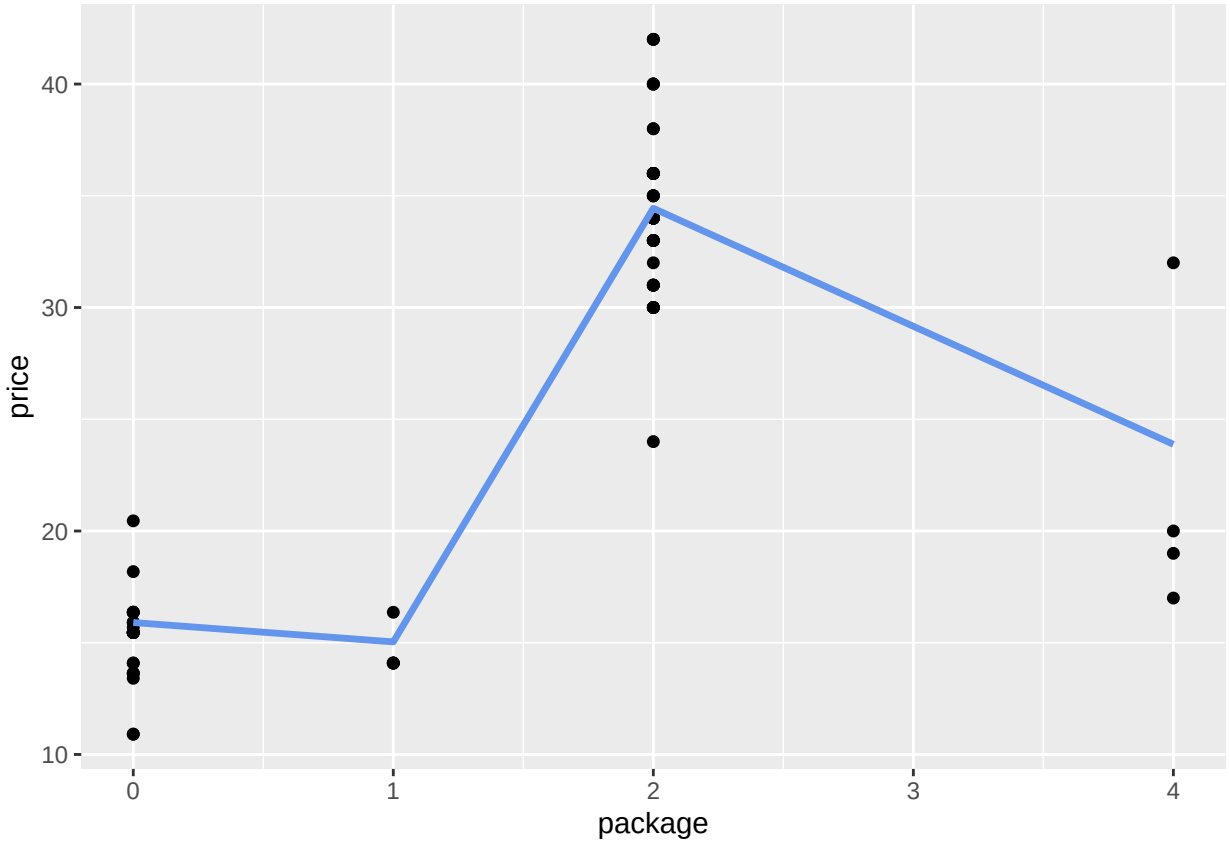
```
# Bind encoded package column to the results
poly_results <- poly_results %>%
  bind_cols(package_encode %>%
  rename(package_integer = package)) %>%
  relocate(package_integer, .after = package)

# Print new results data frame
poly_results %>%
  slice_head(n = 5)
```

```
## # A tibble: 5 x 4
##   package package_integer price .pred
##     <int>           <int> <dbl> <dbl>
## 1       0               0  13.6  15.9
## 2       0               0  16.4  15.9
## 3       0               0  16.4  15.9
## 4       0               0  13.6  15.9
## 5       0               0  15.5  15.9
```

OK, now let's take a look!

Question 4: Create a scatter plot that takes the poly_results and plots package vs. price. Then draw a line showing our model's predicted values (.pred). Hint: you'll need separate geoms for the data points and the prediction line.

```
# Make a scatter plot
poly_results %>%
  ggplot(mapping = aes(x = package, y = price)) +
    geom_point(size = 1.6) +
    geom_line(aes(y = .pred), color = "cornflowerblue", linewidth = 1.2)
```

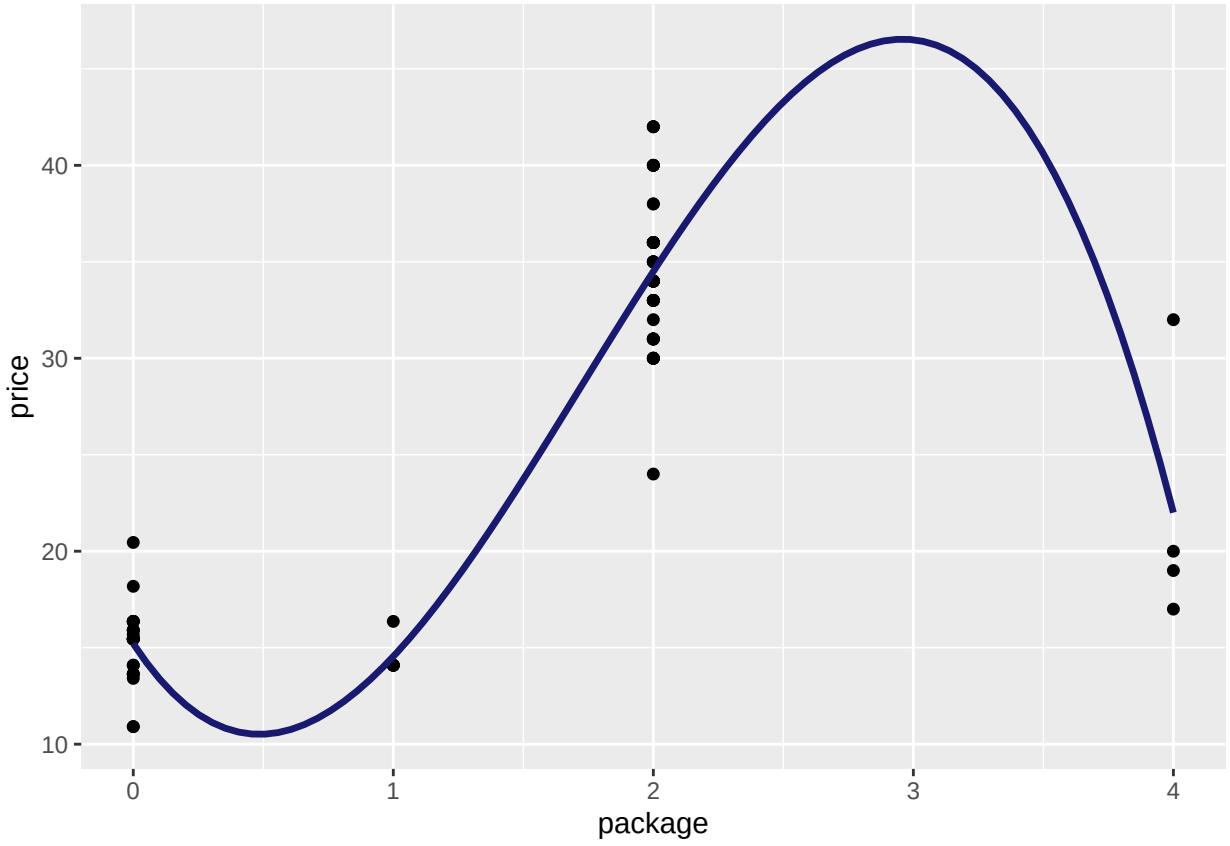


You can see that a curved line fits your data much better.

Question 5: Now make a smoother line by using geom_smooth instead of geom_line and passing it a polynomial formula like this: geom_smooth(method = lm, formula = y ~ poly(x, degree = 3), color = "midnightblue", size = 1.2, se = FALSE)

```
# Make a smoother scatter plot
poly_results %>%
  ggplot(mapping = aes(x = package, y = price)) +
    geom_point(size = 1.6) +
    geom_smooth(method = lm, formula = y ~ poly(x, degree = 3), color = "midnightblue", size = 1.2, se =
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```



OK, now it's your turn to go through the process one more time.

Question 6: Choose a new predictor variable (anything not involving package type) in this dataset.

- **For this model, I will use city name as a predictor variable.**

Question 7: Determine its correlation with the outcome variable (price). (Remember we calculated a correlation matrix last week)

- **The correlation between price and city is 0.32.**

```
# calculate correlation between variety and price
cor(baked_pumpkins$city_name, baked_pumpkins$price)
```

```
## [1] 0.3236397
```

Question 8: Create and test a model for your new predictor: - Create a recipe

```
# create polynomial recipe
city_recipe <- recipe(price ~ city_name, data = pumpkins_train) %>%
  step_integer(all_predictors(), zero_based = TRUE) %>%
  step_poly(all_predictors(), degree = 4)

city_recipe
```

```
##

## -- Recipe ----------------------------------------------------------------------

##

## -- Inputs

## Number of variables by role

## outcome:   1
## predictor: 1

##

## -- Operations

## * Integer encoding for: all_predictors()

## * Orthogonal polynomials on: all_predictors()
```

- Build a model specification (linear or polynomial)

```
# Create a model specification called poly_spec
city_spec <- linear_reg() %>%
  set_engine("lm") %>%
  set_mode("regression")
```

- Bundle the recipe and model specification into a workflow

```
# Hold modeling components in a workflow (polynomial)
city_wf <- workflow() %>%
  add_recipe(city_recipe) %>%
  add_model(city_spec)
```

- Create a model by fitting the workflow

```
# Create polynomial model
city_wf_fit <- city_wf %>%
  fit(data = pumpkins_train)
```

- Evaluate model performance on the test data

```
# Make price predictions on test data
city_results <- city_wf_fit %>%
  predict(new_data = pumpkins_test) %>%
  bind_cols(pumpkins_test %>%
  select(c(city_name, price))) %>%
  relocate(.pred, .after = last_col())

city_results %>%
  slice_head(n = 10)
```

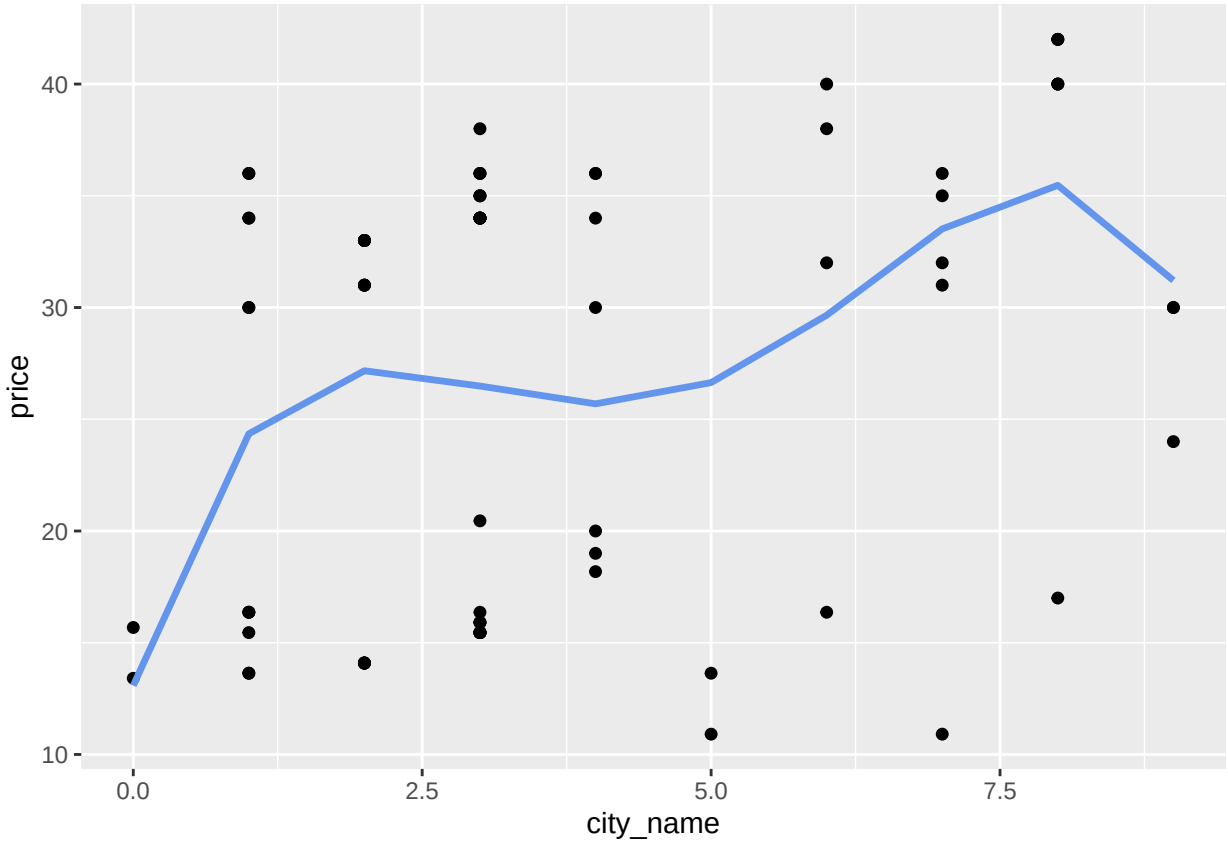
```
## # A tibble: 10 x 3
##    city_name price .pred
##        <int> <dbl> <dbl>
##  1         1  13.6  24.3
##  2         1  16.4  24.3
##  3         1  16.4  24.3
##  4         1  13.6  24.3
##  5         1  15.5  24.3
##  6         1  16.4  24.3
##  7         1  34    24.3
##  8         1  30    24.3
##  9         1  30    24.3
## 10         1  34    24.3
```

```
# look at metrics
metrics(data = city_results, truth = price, estimate = .pred)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       9.15
## 2 rsq     standard       0.142
## 3 mae     standard       8.26
```

- Create a visualization of model performance

```
# Make a scatter plot
city_results %>%
  ggplot(mapping = aes(x = city_name, y = price)) +
    geom_point(size = 1.6) +
    geom_line(aes(y = .pred), color = "cornflowerblue", linewidth = 1.2)
```



Choose Lab 2 due 1/24 at 11:59 PM