Making Inferences about Features

DATA 202 21FA

Logistics

- Final projects
- Midterm 2??
- Discussions
- Quizzes and Homework

What's your goal?

- **Predict** unseen labels
  - How much will this house sell for?
  - Does this child have autism?
  - Is this a positive or negative movie review?
- **Infer** relationships between features and labels
  - How much does home size affect price?
  - Is DNA methylation a marker of autism?
  - Does "sick" indicate a positive or negative review?
- Understand the **causal** effect of interventions
  - How much will building an addition increase the price of my home?
  - Will antioxidants prevent autism?
  - Will cutting this scene make my movie get better reviews?

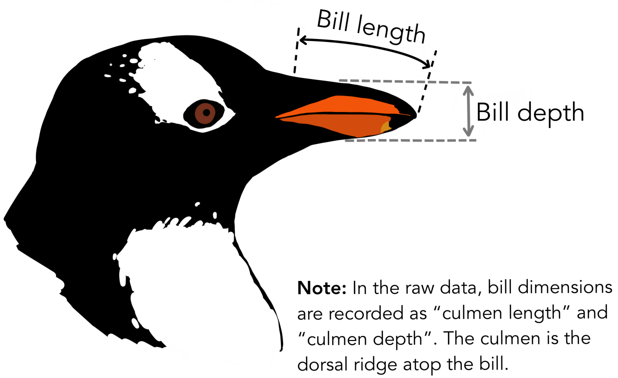
Techniques for Inference

- Classical statistical inference
  - 2-sample t tests, chi squared tests, ANOVA, ...
  - inference about model parameters (coefficient standard errors etc.)
- Variable importance plots
- Benefit of adding each feature

Objectives for Today

- Identify several different approaches for drawing conclusions about features
- Recognize potential challenges in making those inferences

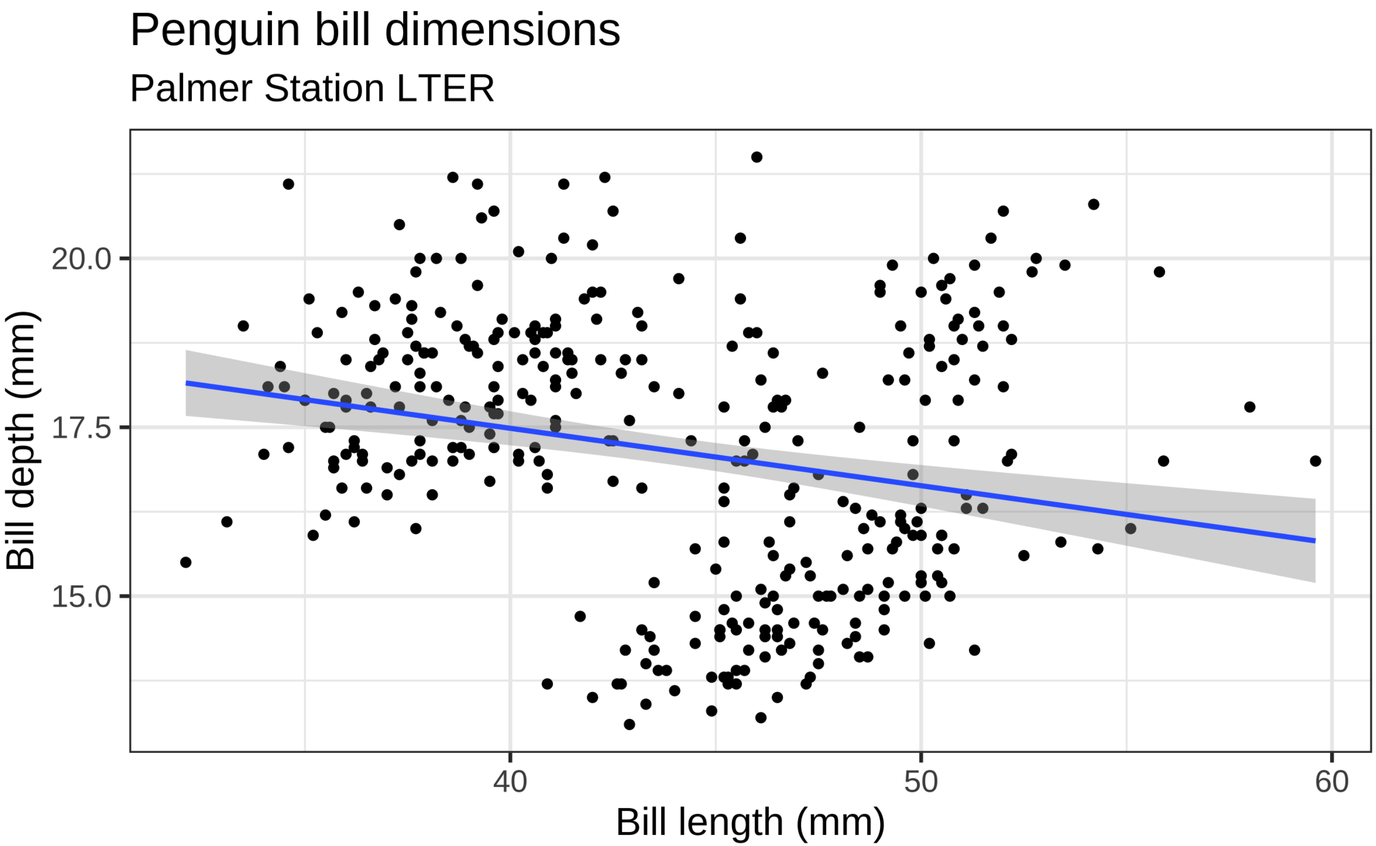
Palmer Penguins



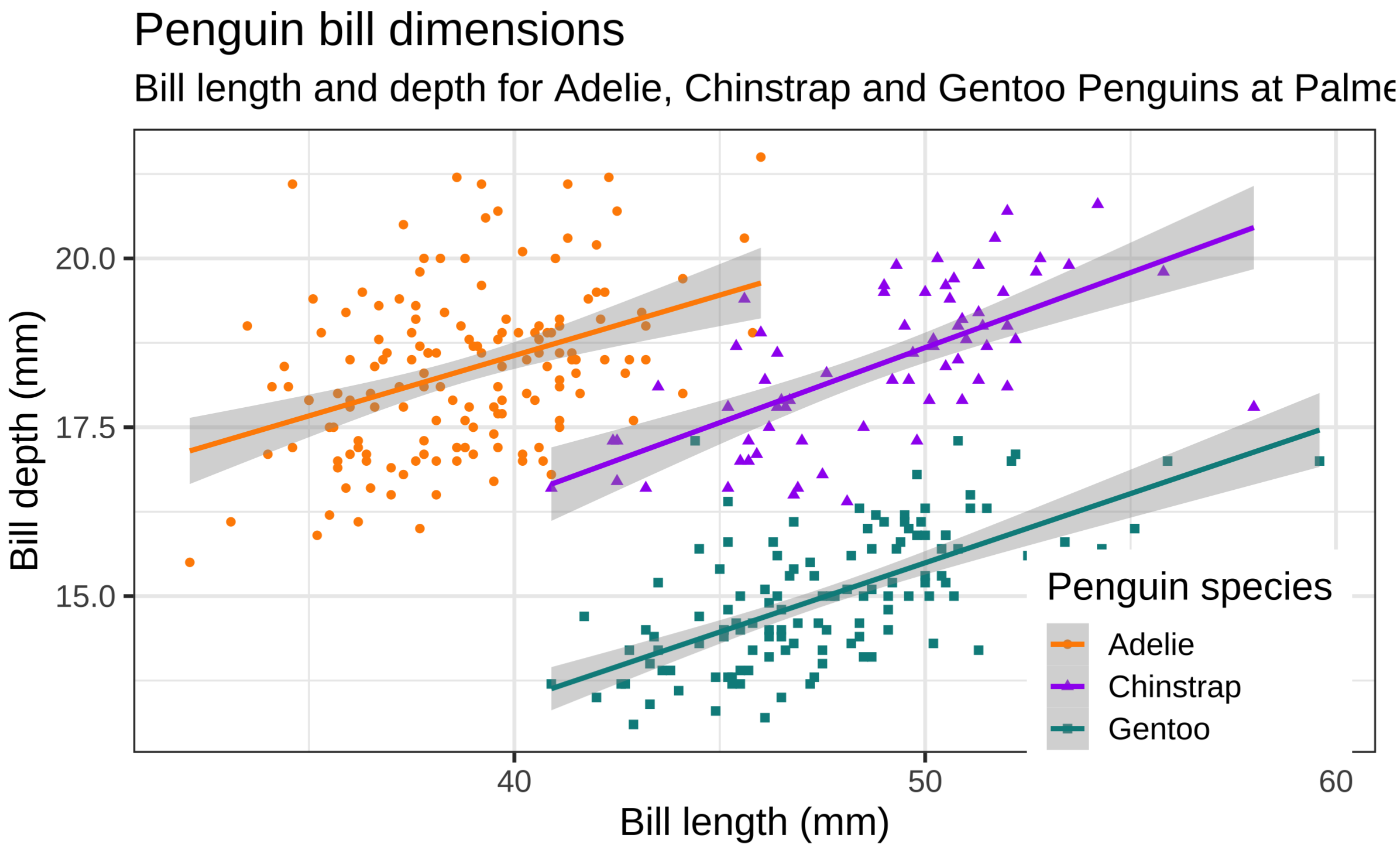

```
library(palmerpenguins)
penguins <- palmerpenguins::pe
  filter(!is.na(bill_length_mm
```

Artwork by @allison_horst

How does bill length relate to bill depth?



How does bill length relate to bill depth?



One big model

Fit Model   Coefficients   Predictions

```
overall_model <- linear_reg() %>%
  fit(bill_depth_mm ~ bill_length_mm, data = penguins)
tidy(overall_model, conf.int = TRUE)
```

```
# A tibble: 2 × 7
  term    estimate std.error statistic  p.value conf.low conf.high
  <chr>      <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Inte…   20.9       0.844      24.7  4.72e-78   19.2      22.5
2 bill_…   -0.0850    0.0191     -4.46 1.12e- 5   -0.123    -0.0475
```

Parallel slopes

Fit Model   Coefficients   Predictions

```
parallel_slopes_model <- linear_reg() %>%
  fit(bill_depth_mm ~ bill_length_mm
      + species,
      data = penguins)
tidy(parallel_slopes_model, conf.int = TRUE)
```

```
# A tibble: 4 × 7
  term    estimate std.error statistic  p.value conf.low conf.high
  <chr>      <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Inte…    10.6      0.683      15.5  2.43e-41    9.25      11.9
2 bill_…     0.200    0.0175     11.4  8.66e-26    0.165      0.234
3 speci…    -1.93     0.224      -8.62 2.55e-16   -2.37      -1.49
4 speci…    -5.11     0.191     -26.7  3.65e-85   -5.48      -4.73
```

Interactions

Fit Model | Coefficients | Predictions

```r
interaction_model <- linear_reg() %>%
  fit(bill_depth_mm ~ bill_length_mm
      * species,
      data = penguins)
tidy(interaction_model, conf.int = TRUE)
```

```
# A tibble: 6 × 7
  term     estimate std.error statistic  p.value conf.low conf.high
  <chr>       <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Inte…    11.4       1.14      10.0   7.28e-21   9.17      13.6  
2 bill_…     0.179     0.0293     6.11  2.76e- 9   0.121      0.236
3 speci…    -3.84      2.05      -1.87  6.24e- 2  -7.88       0.200
4 speci…    -6.16      1.75      -3.51  5.09e- 4  -9.61      -2.71 
5 bill_…     0.0434    0.0456     0.952 3.42e- 1  -0.0463     0.133
6 bill_…     0.0260    0.0405     0.642 5.22e- 1  -0.0537     0.106
```

LINE Assumptions for making inference

- **Linearity**: there's actually a linear relationship
- **Independence**: there's no pattern to the errors
- **Normality of residuals**: no major outliers
- **Equal variance of residuals**: variability doesn't change systematically

These are colloquial wordings. See textbook section E.4 for technical and how to verify.

Examples of broken assumptions

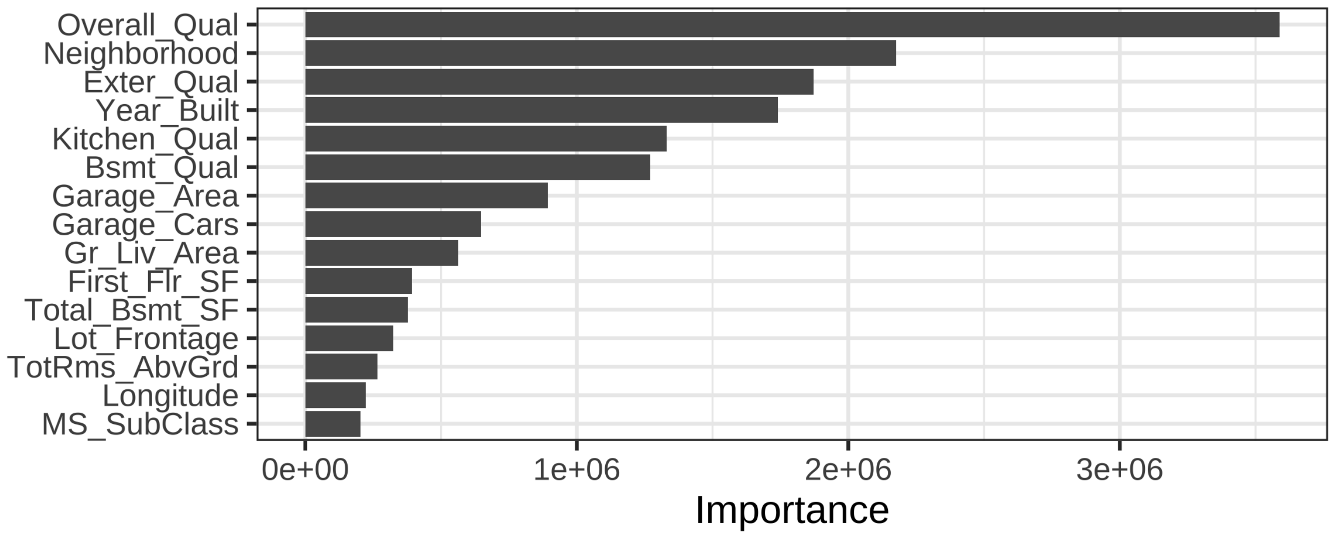
- **Linearity**: the hotter the temperature, the more each degree of temperature matters
- **Independence**: ridership for adjacent hours is similar
- **Normality of residuals**: spike in demand after championship game
- **Equal variance of residuals**: bigger variability in demand on weekends vs weekdays

Variable Importance Plots

```r
regresion_workflow <- workflow() %>%  add_model(decision_tree(mode = "regression") %>% set_e

model <- regresion_workflow %>%
  add_recipe(recipe(Sale_Price ~ ., data = ames_train)) %>%
  fit(data = ames_train)

model %>% extract_fit_engine() %>% vip::vip(num_features = 15L)
```



How much does it help to have a feature in?

```
regression_workflow %>%
  add_recipe(recipe(Sale_Price ~ ., data = ames_train)) %>%
  fit_resamples(resamples = resamples, metrics = metric_set(mae, rmse)) %>% collect_metrics(
```

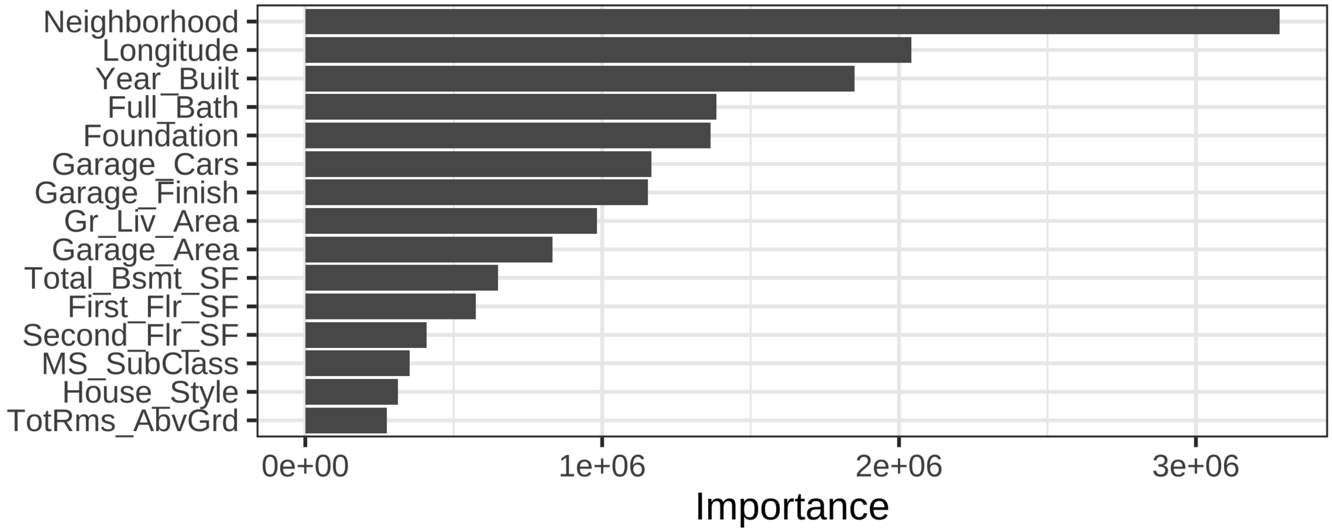
```
# A tibble: 2 × 6
  .metric .estimator  mean     n std_err .config
  <chr>   <chr>      <dbl> <int>   <dbl> <chr>
1 mae     standard    25.7    10   0.438 Preprocessor1_Model1
2 rmse    standard    36.1    10   1.16  Preprocessor1_Model1
```

```
regression_workflow %>%
  add_recipe(recipe(Sale_Price ~ ., data = ames_train) %>% step_rm(ends_with("Qual"))) %>%
  fit_resamples(resamples = resamples, metrics = metric_set(mae, rmse)) %>% collect_metrics(
```

```
# A tibble: 2 × 6
  .metric .estimator  mean     n std_err .config
  <chr>   <chr>      <dbl> <int>   <dbl> <chr>
1 mae     standard    25.6    10   0.627 Preprocessor1_Model1
2 rmse    standard    36.5    10   1.10  Preprocessor1_Model1
```

Variable Importance without the Quality Features

```
regresion_workflow %>%
  add_recipe(recipe(Sale_Price ~ ., data = ames_train) %>% step_
  fit(data = ames_train) %>%
  extract_fit_engine() %>% vip::vip(num_features = 15L)
```



Appendix: code

```
include_graphics("https://raw.githubusercontent.com/allisonhorst/palmerpenguins/master/man/f
include_graphics("https://raw.githubusercontent.com/allisonhorst/palmerpenguins/master/man/f
ggplot(penguins, aes(x = bill_length_mm, y = bill_depth_mm)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Penguin bill dimensions", subtitle = "Palmer Station LTER", x = "Bill length
ggplot(penguins, aes(x = bill_length_mm, y = bill_depth_mm, color = species, shape = species
  geom_point() +
  geom_smooth(method = "lm") +
  scale_color_manual(values = c("darkorange","purple","cyan4")) +
  labs(title = "Penguin bill dimensions",
       subtitle = "Bill length and depth for Adelie, Chinstrap and Gentoo Penguins at Palmer
       x = "Bill length (mm)",
       y = "Bill depth (mm)",
       color = "Penguin species",
       shape = "Penguin species") +
  theme(legend.position = c(0.85, 0.15),
        legend.background = element_rect(fill = "white", color = NA))
#data(ames, package = "modeldata")
ames <- AmesHousing::make_ames()
ames_all <- ames %>%
  filter(Gr_Liv_Area < 4000, Sale_Condition == "Normal") %>%
  mutate(across(where(is.integer), as.double)) %>%
  mutate(Sale_Price = Sale_Price / 1000)
rm(ames)
set.seed(10) # Seed the random number generator
ames_split <- initial_split(ames_all, prop = 2 / 3)
ames_train <- training(ames_split)
ames_test <- testing(ames_split)
set.seed(0)
resamples <- vfold_cv(ames_train, v = 10)
```