IMS 8 & 25: Linear regression with multiple predictors

Packages

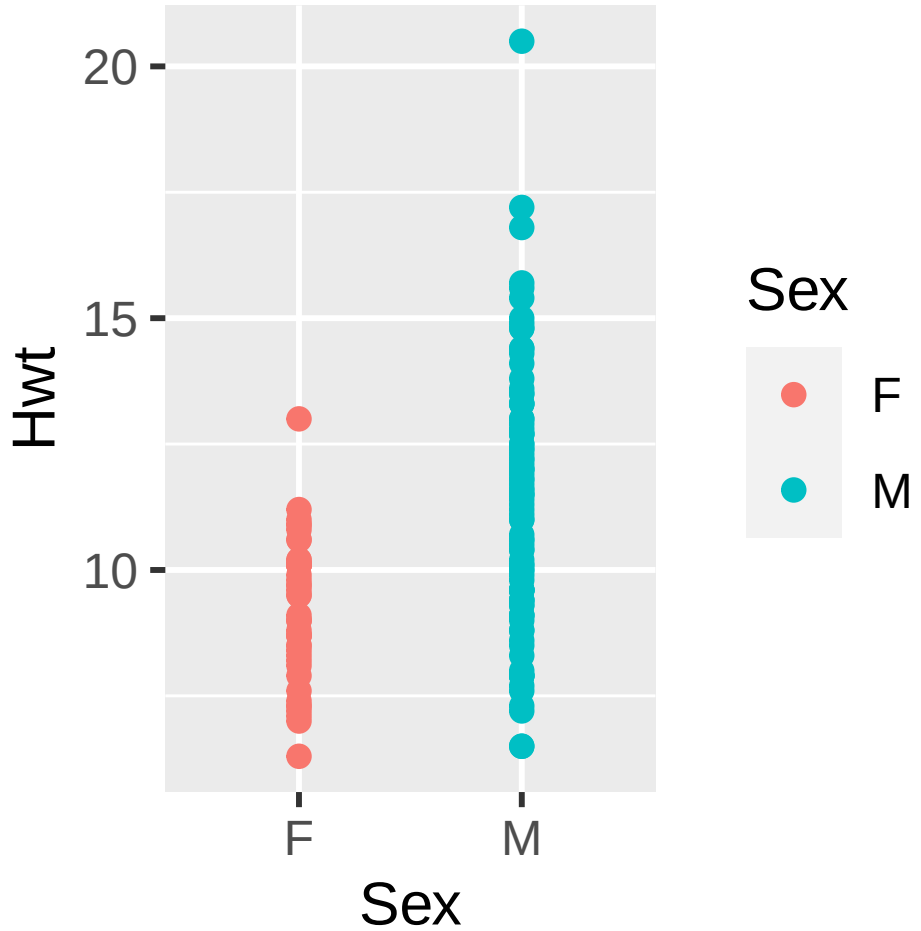
```
library(MASS)       # for cats data set
library(tidyverse) # for ggplot functions/plotting
library(broom)      # for tidy() function
library(openintro) # for loan data set
library(GGally)     # for ggpairs() function
library(palmerpenguins) # penguin data set
```

Example

- Larger heart weights indicate a higher risk of heart attacks/disease in cats; however, heart weight is hard to measure.
- Want to see if there is a relationship between heart weight (`Hwt`) and sex (`Sex`) for domestic cats.
- If so, we will have a better idea of which cats are at risk for heart attacks/disease.

Example

Is sex a good predictor variable for the heart weight?



Categorical Variable with Two Levels

Randomly pick 6 rows to preview:

```
set.seed(62)
index <- sample(1:nrow(cats), 6)
cats[index, ]
```

```
    Sex Bwt  Hwt
69    M 2.5  9.3
133   M 3.5 15.6
29    F 2.3 10.6
84    M 2.7 10.4
4     F 2.1  7.2
135   M 3.5 17.2
```

The variable `Sex` is a categorical variable with two levels: `M` and `F`

Example

```
lm(Hwt ~ Sex, data = cats) |>
  tidy(conf.int = T, conf.level = .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)     9.20     0.325      28.3  2.96e-60     8.56      9.84
2 SexM            2.12     0.396       5.35 3.38e- 7     1.34      2.90
```

Example

Hwt for female group

```
# 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)     9.20     0.325      28.3  2.96e-60     8.56      9.84
2 SexM            2.12     0.396       5.35 3.38e- 7     1.34      2.90
```

The change in Hwt due to being male, compared to female group

Categorical Variable with Two Levels

Interpretation for $\hat{\beta}_0, \hat{\beta}_1$ with an indicator variable

- The expected mean value of Y for a subject in the level-0 group is $\hat{\beta}_0$
- The expected mean value of Y changes by $\hat{\beta}_1$ units when a subject is in level-1 group in comparison to the level-0 group

Example:

- The expected mean value of `Hwt` when `Sex = F` is 9.20
- We expect the mean value of `Hwt` to increases by 2.12 grams when `Sex = M` in comparison to the `Sex = F` group (i.e. The expected mean value of `Hwt` when `Sex = M` is $9.20 + 2.12 = 11.32$)

Categorical Variable with Multiple Categories

We will consider data about loans from the peer-to-peer lender, Lending Club. The outcome variable we would like to better understand is the **interest rate** assigned to the loan.

The dataset includes results from 10,000 loans, and we'll be looking **verified income** as the predictor variable.

```
  interest_rate verified_income
1         14.07         Verified
2         12.61     Not Verified
3         17.09 Source Verified
4          6.72     Not Verified
5         14.07         Verified
6          6.72     Not Verified
```

Categorical Variable with Multiple Categories

```
lm(interest_rate ~ verified_income, data = loans_full_schema) |>
  tidy()
```

```
# A tibble: 3 x 5
  term                              estimate std.error statistic   p.value
  <chr>                                <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)                          11.1     0.0809     137.   0
2 verified_incomeSource Verified        1.42    0.111       12.8 3.79e- 37
3 verified_incomeVerified               3.25    0.130       25.1 8.61e-135
```

Categorical Variable with Multiple Categories

Interest Rate for "Not Verified" group.

```
# A tibble: 3 x 5
  term                              estimate std.error statistic   p.value
  <chr>                                <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)                          11.1     0.0809     137.   0
2 verified_incomeSource Verified        1.42    0.111       12.8 3.79e- 37
3 verified_incomeVerified               3.25    0.130       25.1 8.61e-135
```

Change in Interest Rate for "Source Verified" group, in comparison to "Not Verified" group.

Change in Interest Rate for "Verified" group, in comparison to "Not Verified" group.

Categorical Variable with Multiple Categories

- The missing level is called the **reference level** and it represents the default level that other levels are measured against.
- A categorical variable that has K levels where $K > 2$, software will provide a coefficient for $K - 1$ of those levels.

Discussion Questions

- What information is useful when determining an interest rate on a loan?
- Why would "verified income" be a useful variable to look into? Why not use "income" instead?
- What if we considered other variables like criminal history?

Proxy Variables

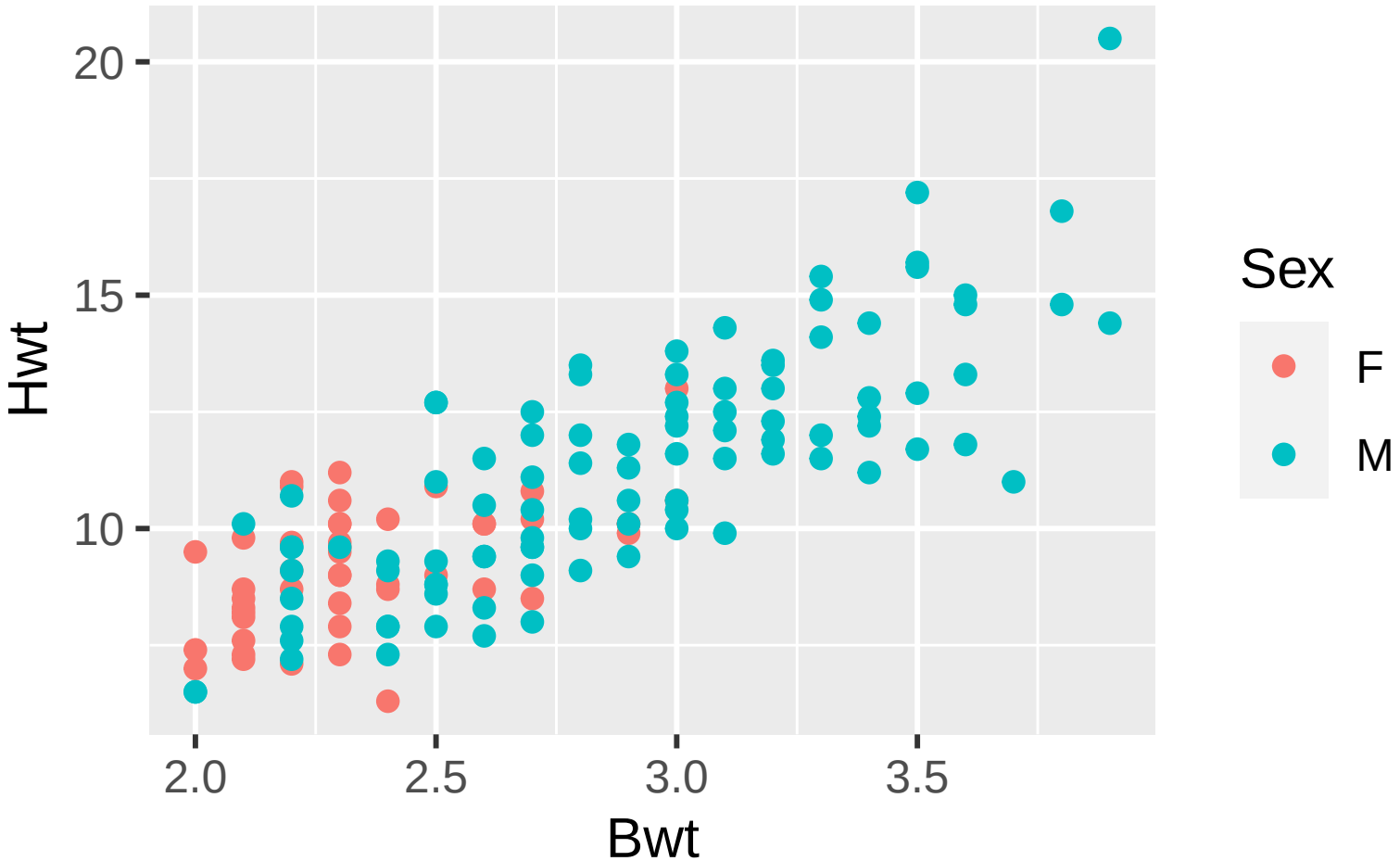
- A **proxy variable** is a variable that is not in itself directly relevant, but that serves in place of an unobservable or immeasurable variable.
- Be very careful about interpretation! We use proxy variables often.

Linear Regression with a Multiple Predictors

- We are not limited to only one predictor variable.
- We can add multiple predictors to models.
- Want to see if sex (`Sex`) and body wieght (`Bwt`) can be used to predict heart weight (`Hwt`) for domestic cats.

Linear Regression with a Multiple Predictors

```
ggplot(cats, aes(x = Bwt, y = Hwt, color = Sex))+
  geom_point()
```



Linear Regression with a Multiple Predictors

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

- x_1: sex, `M` for male or `F` for female.
- x_2: body weight (`Bwt`) in kilograms
- y: heart weight (`Hwt`) in grams
- e: error

Linear Regression with Multiple Predictors

```
lm(Hwt ~ Sex + Bwt, data = cats) |>
  tidy()
```

```
# A tibble: 3 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  -0.415      0.727    -0.571 5.69e- 1
2 SexM         -0.0821     0.304    -0.270 7.88e- 1
3 Bwt           4.08       0.295    13.8   5.12e-28
```

Linear Regression with Multiple Predictors

Multiple predictors allows us to summarize the effect while controlling for a variable. For example,

$$y = -0.415 - 0.0832x_1 + 4.08x_2$$

the coefficient of x_1 is $\widehat{\beta}_1 = -0.0832$ regardless if $x_2 = 2$, 2.7, or 3, etc.

Linear Regression with Multiple Predictors

Hwt for female cats with Bwt =0

```
# A tibble: 3 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  -0.415       0.727    -0.571 5.69e- 1
2 SexM         -0.0821      0.304    -0.270 7.88e- 1
3 Bwt           4.08        0.295    13.8   5.12e-28
```

Change in Hwt when Sex = M, holding Bwt constant.

Change in Hwt for Bwt, holding gender constant

Linear Regression with Multiple Predictors

Interpretation:

- The expected mean value of `Hwt` when `Sex = F` and `Bwt =0` is -0.415.
- We expect the mean value of `Hwt` to decreases by -0.0821 grams when `Sex = M` in comparison to the `Sex = F` group, holding `Bwt` constant.
- For every 1 kilogram increase in `Bwt` we expect the mean value of `Hwt` to increase by 4.08 grams, holding `Sex` constant.

Practice Problem

Let x_1: sex; x_2: body weight (`Bwt`) in kilograms; y: heart weight (`Hwt`) in grams. Recreate the three different models we explored with the `cats` data.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

$$y = \beta_0 + \beta_1 x_1 + e$$

$$y = \beta_0 + \beta_2 x_2 + e$$

What do you notice about the three different models? What do you wonder?

Inference with Multiple Variables

Before, with a single predictor variable:

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

Now, with more than one predictor variable:

$$H_0 : \beta_i = 0, \text{given other variables in the model}$$

$$H_A : \beta_i \neq 0, \text{given other variables in the model}$$

A low p-value (or CI that do not contain 0) tell us that a variable acts as an important predictor in the model, even when controlling the effects of other variables.

Inference with Multiple Variables

```
lm(Hwt ~ Sex + Bwt, data = cats) |>
  tidy(conf.int = T, conf.level = 0.95)
```

```
# A tibble: 3 x 7
  term        estimate std.error statistic  p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)  -0.415       0.727    -0.571 5.69e- 1   -1.85      1.02
2 SexM         -0.0821      0.304    -0.270 7.88e- 1   -0.683     0.519
3 Bwt           4.08        0.295    13.8   5.12e-28    3.49      4.66
```

P-values

Conf. Intervals

Recall: R^2, Coefficient of Determination

In Chapter 7 we learned about the **Coefficient of Determination** (R^2), which measures the proportion of the variation in the outcome variable Y that is explained by the linear regression model with single predictor X

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

It is the square of the correlation coefficient.

R^2_{adj}, Adjusted Coefficient of Determination

For linear regression models with multiple predictors we use **Adjusted Coefficient of Determination** (R^2_{adj}). Let n be the number of observations, and p is the number of βs (not including β_0)

$$R^2_{adj} = 1 - \frac{SSE}{SST}\left(\frac{n-1}{n-p-1}\right)$$

Measures the proportion of the variation in the outcome variable Y that is explained by the linear regression model with all of the predictors we used.

R^2_{adj}, Adjusted Coefficient of Determination

```
fit <- lm(Hwt ~ Sex + Bwt, data = cats)
summary(fit)
```

```
Call:
lm(formula = Hwt ~ Sex + Bwt, data = cats)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5833 -0.9700 -0.0948  1.0432  5.1016

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.4149     0.7273  -0.571    0.569
SexM         -0.0821     0.3040  -0.270    0.788
Bwt           4.0758     0.2948  13.826   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.457 on 141 degrees of freedom
Multiple R-squared:  0.6468,    Adjusted R-squared:  0.6418
F-statistic: 129.1 on 2 and 141 DF,  p-value: < 2.2e-16
```

Unadjusted Version

Adjusted Version

Practice

Let x_1: sex; x_2: body weight (`Bwt`) in kilograms; y: heart weight (`Hwt`) in grams. Compare R^2_{adj} for the three different models we explored with the `cats` data.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

$$y = \beta_0 + \beta_1 x_1 + e$$

$$y = \beta_0 + \beta_2 x_2 + e$$

What do you notice about the different R^2_{adj} and R^2 values? What do you wonder?

Multicollinearity

- Sometimes a set of predictor variables can impact the model in unusual ways, often due to the predictor variables themselves being correlated.
- **Multicollinearity** happens when the predictor variables are correlated within themselves.
- When the predictor variables themselves are correlated, the coefficients in a multiple regression model can be difficult to interpret.
- Check for multicollinearity by looking at the correlation between predictor variables or linear trends between predictor variables.

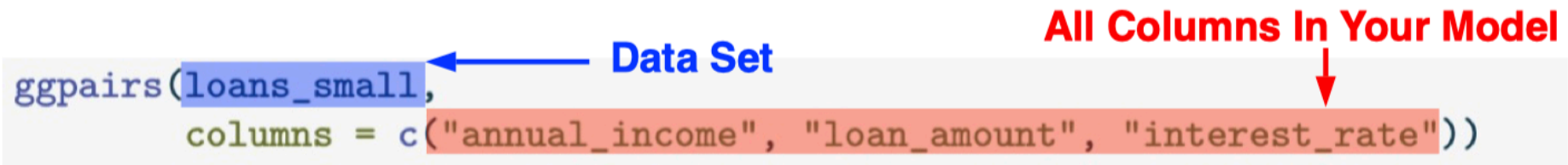
Example - Multicollinearity

We will consider data about loans from the peer-to-peer lender, Lending Club. The outcome variable we would like to better understand is the **interest rate** assigned to the loan.

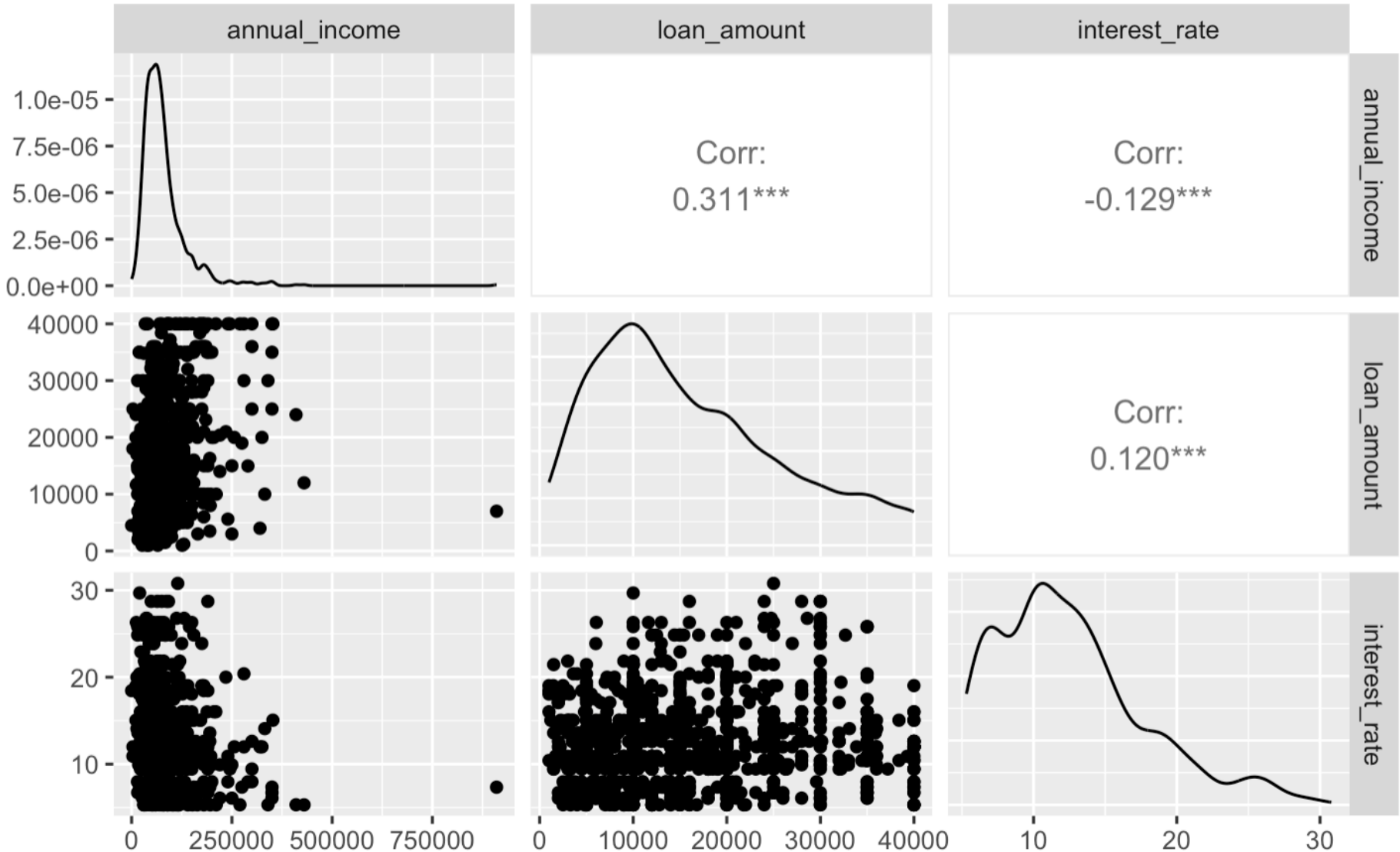
The original `loans_full_schema` dataset includes results from 10,000 loans. We will take a smaller subset of 1000 for simplicity.

```
set.seed(62)
index <- sample(1:nrow(loans_full_schema), 1000)
loans_small <- loans_full_schema[index, ]
```

Check for Multicollinearity



Check for Multicollinearity



Building Models

Things to consider when building models:

- Keep models parsimonious (simple)
- Avoid having two very similar variables (multicollinearity)
- It is okay to have variables that are not statistically significant, they might be important to the research question anyways.
- It is okay to have variables in the model not important to the research question, but you want to control for their effect.

Example

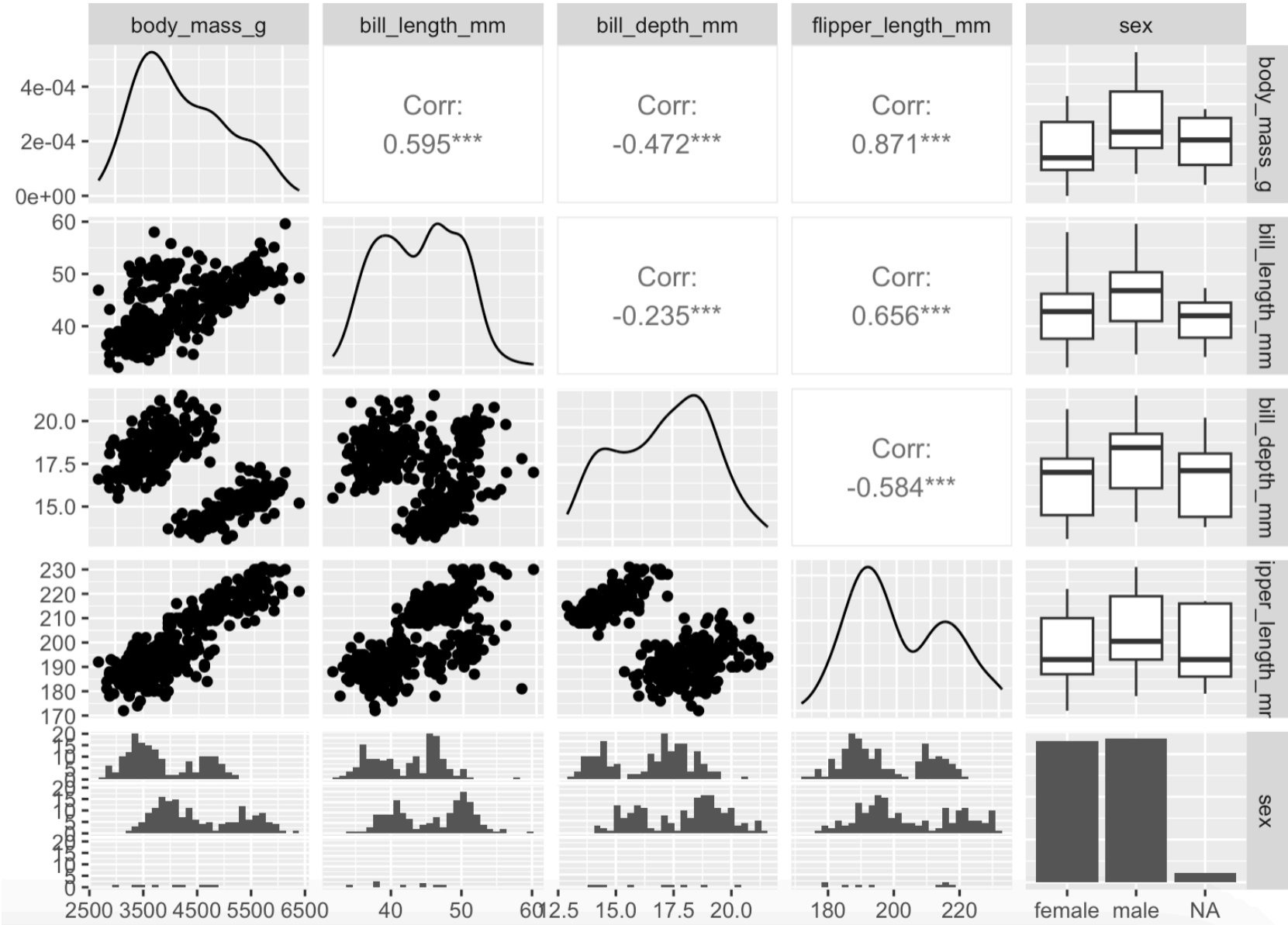
Researchers studying a community of Antarctic penguins collected body measurement (bill length, bill depth, and flipper length measured in millimeters and body mass, measured in grams), and sex (female or male) data on 344 penguins. The data set is called `penguin` in R.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2288.4650	631.5802	-3.6234	0.0003
bill_length_mm	-2.3287	4.6843	-0.4971	0.6194
bill_depth_mm	-86.0882	15.5698	-5.5292	0.0000
flipper_length_mm	38.8258	2.4478	15.8618	0.0000
sexmale	541.0285	51.7098	10.4628	0.0000

a) Calculate the residual for a male penguin that weighs 3750 grams with the following body measurements: `bill_length_mm` $= 39.1$, `bill_depth_mm` $= 18.7$, `flipper_length_mm` $= 181$. Does the model overpredict or underpredict this penguin's weight?

b) Interpret the slope for `bill_depth_mm` in context.

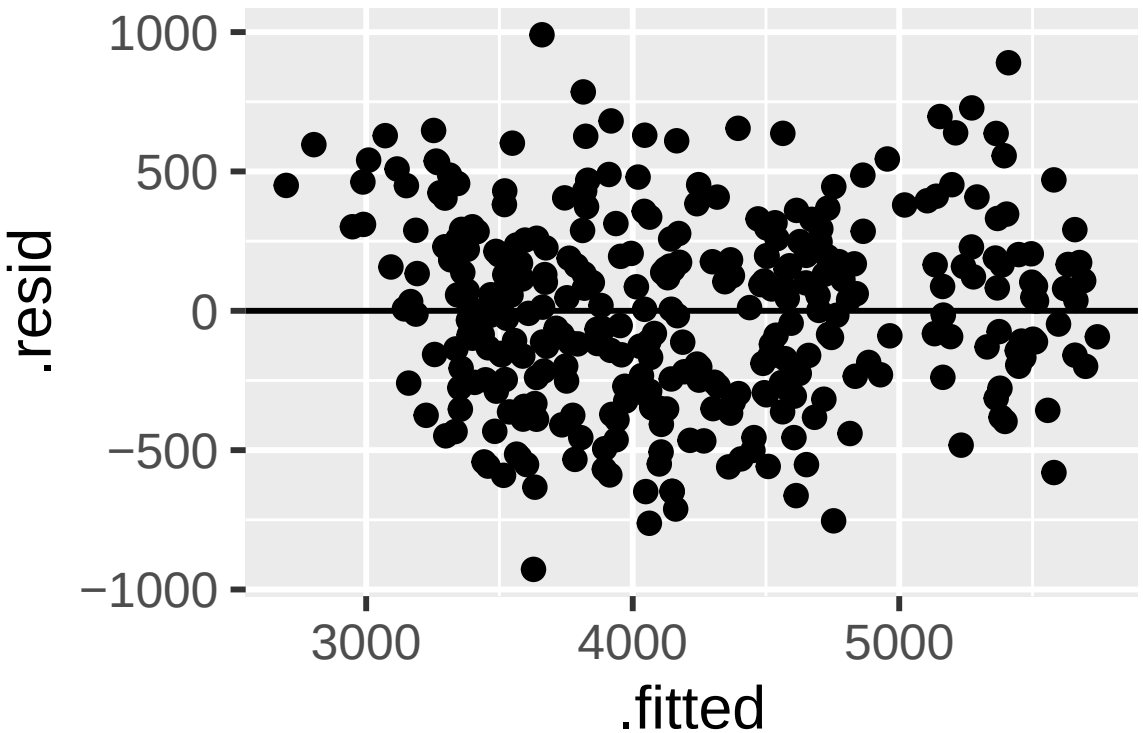
Example


body_mass_g
bill_length_mm
bill_depth_mm
flipper_length_mm
sex
Corr:
0.595***
Corr:
-0.472***
Corr:
0.871***
Corr:
-0.235***
Corr:
0.656***
Corr:
-0.584***
4e-04
2e-04
0e+00
60
50
40
20.0
17.5
15.0
230
220
210
200
190
180
170
2500 3500 4500 5500 6500
40
50
60
12.5
15.0
17.5
20.0
180
200
220
female
male
NA

Example

c) There are six pairs of continuous variables described in the figure, making six different scatter plots. Rate the pairwise relationships from most correlated to least correlated.

d) The residual plot is provided below. The (unadjusted) R^2 is 0.823, and R^2_{adj} is 0.8208. Does the model we created appear to be a good fit?



e) Recreate all plots and results in this problem in R.