STA130 W11

Monday, November 25, 2019 2:12 PM

- Correlation (r) to quantify linear association between two numerical variables
- Simple linear regression model

Strength and direction

Response — Intercept, slope, random error term

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

numerical — Numerical predictor

- Fitted regression line (estimated regression coefficients from `lm()` or closed form expressions)

Estimated regression coefficients from sample data

(X1,y1),…,(x2,y2)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Coefficient of determination R^2

$$R^2 = \frac{\text{"Explained" Variation}}{\text{Total Variation}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- RMSE to assess predictive performance of a linear regression model for prediction

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$$

2 / 40

R^2	$RMSE$
$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Range: 0 - 1	Range: 0 to ∞
Large value $\Rightarrow$ good fit	Small value $\Rightarrow$ good fit
No units	Same units as y
Tells us proportion of variation in y which is explained by the model	Used to assess prediction error (with training or test data)

*R^2 only makes sense to calculate for the data used to build the model
*RMSE: compare RMSE for training and testing data

Hypothesis tests to determine if there is a **real** linear association between two variables
Linear regression with a categorical predictor
Multiple linear regression models (i.e. regression with more than one predictor)

- Model fitting Comparing predictive models

Predicting Body Mass Index (BMI) from waist circumference

Body measurements for 507 physically active adults

```
body <- read.csv("bodydat.csv")
body %>% select(age, weight, height, gender, waist) %>% glimpse()
```

```
## Observations: 507
## Variables: 5
## $ age    <int> 21, 23, 28, 23, 22, 21, 26, 27, 23, 21, 23, 22, 20, 26,...
## $ weight <dbl> 65.6, 71.8, 80.7, 72.6, 78.8, 74.8, 86.4, 78.4, 62.0, 8...
## $ height <dbl> 174.0, 175.3, 193.5, 186.5, 187.2, 181.5, 184.0, 184.5,...
## $ gender <fct> Male, Male, Male, Male, Male, Male, Male, Male, Male, M...
## $ waist  <dbl> 71.5, 79.0, 83.2, 77.8, 80.0, 82.5, 82.0, 76.8, 68.5, 7...
```

- lengths in cm
- weight in kg

$$\text{BMI} = \frac{\text{weight (in kg)}}{[\text{height (in meters)}]^2}$$

```
body <- body %>% mutate(BMI = weight / (height/100)^2)
```

Y x

```
summary(lm(BMI ~ waist, data=body))$coefficients
##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 4.6566078 0.585776319  7.949464  1.233300e-14
## waist       0.2443089 0.007532967 32.431966 1.495616e-125
summary(lm(BMI ~ waist, data=body))$r.squared
## [1] 0.6756234
```

Approximately 68% of the values of BMI can be predicted from waist circumference

Wrong - 68% of the variable in BMI is explained by **own** model
Q1 - What would the regression line look like if there was no association between waist circumference and BMI?
Horizontal line with slope of 0 (i.e. β1-hat = 0)
Q2 - How can we determine whether the association we observed is "real" or just due to sampling variability?
Significant test(hypothesis test!)

Inference for simple linear regression

Population level parameter

$$\text{Model:} \qquad y = \beta_0 + \beta_1 x_i + \epsilon_i$$

where y_i and x_i are the BMI and waist circumference of individual i, repspectively, and ϵ_i is the difference between individual i's response and the mean BMI for people with the same waist circumference as them.

Based on this model, how can we write hypotheses to test if there is an association between waist circumference and BMI in physically-active adults?

H0 = β1 = 0 (there is no linear regression between x and y)
Ha = β1 ≠ 0

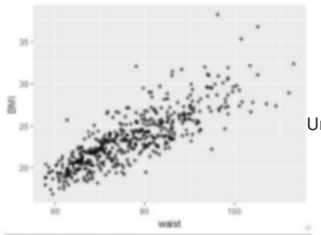
The inference results given by `lm()` are based on the **Students' t distribution** (a continuous probability distribution) so rely on some assumptions:

A1: Linear pattern between x and y

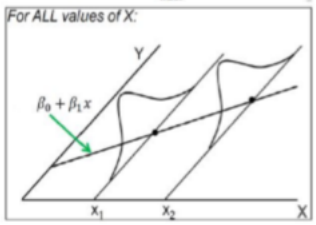
A2: Constant variance in y for all values of x

A3: Independent observations

A4: Residuals follow a normal (a bell-shaped continuous probability distribution)



Unequal variance might looks like



We checked that these conditions are reasonable for these data so it is valid to make inferences based on the p-values generated by R for this regression.

You'll learn more about strategies to check these assumptions and to deal with violations in future statistics courses

If one or more of the assumptions above are not reasonable, then the inference results given by `lm()` may not be valid, although the model may still be used for prediction.

10 / 40

```
summary(lm(BMI ~ waist, data=body))$coefficients
##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 4.6566078 0.585776319  7.949464  1.233300e-14 (beta0)
## waist       0.2443089 0.007532967 32.431966 1.495616e-125(beta1)
```

R gives us p-values for hypothesis tests of the form
Ho: beta = 0 Ha: beta ≠0
Do we have evidence against the null hypothesis that the slope is different from 0 (i.e. the null hypothesis that there is no association between waist circumference and BMI)? If yes, how strong is this evidence?
P-value : 1.49x10^-125 ~= 0
Very strong evidence against the null hypothesis of no linear association between waist and BMI

What other factors might also affect BMI?
There are many other possible predictors in our dataset
Let's see if the distribution of BMI varies for men and women?

Regression with a categorial predictor

$$\text{Model:} \qquad y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where y_i is the BMI for individual i, ϵ_i is the random error term for individual i, and

$$x_i = I(\text{individual i is male}) = \begin{cases} 1 & \text{if individual i is male} \\ 0 & \text{if individual i is female} \end{cases}$$

Gender is a categorical predictor with levels "male" and "female", but we need to convert these levels to numbers

- We encode categorical predictors as **indicator variables** (also called **dummy variables**)
- We need to pick a **baseline value** (i.e. the level corresponding to $x = 0$); here `female` is the baseline

In the R output, the *genderMale* indicates that "Male" is **not** the baseline level: the level which is missing from the output is the baseline!
summary(lm(BMI ~ gender, data=body))$coefficients
Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.27793 0.1886217 118.109055 0.000000e+00
genderMale 2.43330 0.2702385 9.004269 4.435794e-18
Male is no the base line here
The fitted regression line is y-hati = 22.3 + 2.4 xi ,where is the BMI and
Xi =I(individual is male) .
**x is either 0 or 1, so this really only gives us 2 predictions, one for men and one for women

Men: y= betao + beta1(1) + E
Women: y = betao + beta1(0) +E
H0: beta1 =0
Ha : beta1 ≠ 0
->very strong evidence against the ho

What other method (which we covered in this course) could we also use to investigate whether there is a difference in mean BMI for men and women?
- Randomization test, by shuffling the group labels (M F) to sim. Values under Ho

Model 1:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

BMI waist I(i is male)

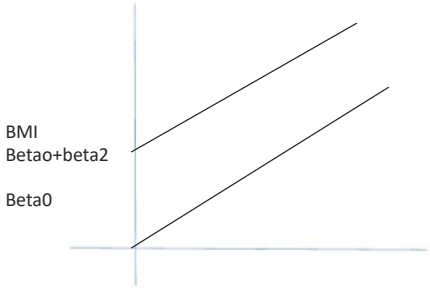
where y_i and x_{1i} are the BMI and waist cirumference for individual i and $x_{2i} = \text{I(individual i is male)}$.

Fitted line for women

Y-hati = beta0-hat + beta1-hatxi + beta2-hat(0)

Fitted line for men

Y-hati = beta0-hat + beta1-hatxi + beta2-hat(1)
=(betao-hat + beta2-hat) + beta1-hatxi



Fitted model -> for multiple linear regression
(y~ x1 + x2)
parallel_lines <- lm(BMI ~ waist + gender, data = body)
summary(parallel_lines)$coefficients
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.7454759 0.651078332 1.144987 2.527582e-01
waist 0.3084726 0.009205015 33.511364 2.358772e-130
genderMale -2.1104205 0.202610484 -10.416146 3.845116e-23

Equation:
Y-hat = 0.745 + 0.308 x (waist circumference) - 2.11 x I(individual is male)

Plotting parallel lines
The augment function (in the library broom) creates a data frame with predicted values (.fitted), residuals, etc ...
library(broom) augment(parallel_lines)
A tibble: 507 x 10
BMI waist gender .fitted .se.fit .resid .hat .sigma .cooksd

```
##  * <dbl> <dbl> <fct>   <dbl>  <dbl> <dbl>  <dbl> <dbl>  <dbl>
##  1  21.7  71.5 Male     20.7  0.161  0.976 0.00906  1.70 1.02e-3
...
```

Join up the fitted values to plot the parallel lines model

```
body %>% ggplot(aes(x=waist, y=BMI, color=gender)) +
    geom_point(alpha=0.5) +
    geom_line(data=augment(parallel_lines),
        aes(y=.fitted, colour=gender), lwd=1.5)
```

Scatterplot

To add the fitted regression line

Line width

Allowing for non-parallel lines

In Model 1, we assumed that the association between waist and BMI was the same for men and women (i.e. same slope)
Let's look to see if gender **modifies the relationship** between waist and BMI - to do this, we add a new independent variable to the model which is the product of *waist* and *gender*: this is an **interaction** term

Model 2:

Interaction term

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$$

Same as before

where y_i and x_{1i} are the BMI and waist cirumference for individual i and $x_{2i} = \mathrm{I}(\text{individual i is male})$.

To fit a linear model with an interaction term in R, use * instead of + between the predictors

```
summary(lm(BMI ~ waist * gender, data = body))$coefficients
##                  Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     -3.4908232 0.94161575 -3.707269  2.327883e-04
## waist            0.3691615 0.01341083 27.527117 1.985004e-102
## genderMale       6.1578438 1.38114262  4.458514  1.018556e-05
## waist:genderMale -0.1083858 0.01792202 -6.047631  2.869562e-09
```

B-hat0
1
2
3

-> coefficient for the interaction between waist cir. And gender

****row names are important!**

Fitted line for women(x2=0)
Y-hat = b0-hat + b1-hatx1 + b2-hat(0) + b3-hatx1(0)
Fitted line for men
Y-hat = b0-hat + b1-hatx1 + b2-hat(1) + b3-hatx1(1)
= (b0-hat + b2-hat) + (b1-hat + b3-hat)x1

We can then plug in values of b-hat from the table of the top to make a prediction for a new observations

Plot of non-parallel lines

```
body %>% ggplot(aes(x=waist, y=BMI, color=gender)) +
 geom_point(alpha=0.5) + geom_smooth(method="lm", se=FALSE)
```

Different intercepts and slopes for men and women

Consider our four regression models

Model A:

$$y_i = \beta_0 + \beta_1 \times \text{waist}_i + \epsilon_i$$

Model B:

$$y_i = \beta_0 + \beta_1 \times \mathrm{I}(\text{individual i is male}) + \epsilon_i$$

Model C:

$$y_i = \beta_0 + \beta_1 \times \text{waist}_i + \beta_2 \times \mathrm{I}(\text{individual i is male}) + \epsilon_i$$

Model D:

$$y_i = \beta_0 + \beta_1 \times \text{waist}_i + \beta_2 \times \mathrm{I}(\text{i is male}) + \beta_3 \times \text{waist}_i \cdot \mathrm{I}(\text{i is male}) + \epsilon_i$$

Comparing our models: Which of our models performs best to predict BMI?
S t r a t e g y
1. Divide sample into training dataset and testing dataset
2. Fit a prediction model (today: a linear regression model) using the training data
3. Make predictions for the testing data and compare the predictions to the true responses

```
set.seed(1210);
n <- nrow(body)
```

```
training_indices <- sample(1:n, size=round(0.8*n))
train <- body[training_indices,];
y_train <- train$BMI;

# T e s t i n g d a t a s e t i n c l u d e s a l l o b s e r v a t i o n s N O T i n t h e t r a i n i n g d a t a
test <- body[-training_indices,];
y_test <- test$BMI;

# F i t m o d e l s t o t r a i n i n g d a t a
modA_train <- lm(BMI ~ waist, data=train)
modB_train <- lm(BMI ~ gender, data=train)
modC_train <- lm(BMI ~ waist + gender, data=train)
modD_train <- lm(BMI ~ waist * gender, data=train)

# M a k e p r e d i c t i o n s f o r t e s t i n g d a t a u s i n g t r a i n i n g m o d e l
yhat_modA_test <- predict(modA_train, newdata = test)
yhat_modB_test <- predict(modB_train, newdata = test)
yhat_modC_test <- predict(modC_train, newdata = test)
yhat_modD_test <- predict(modD_train, newdata = test)

# M a k e p r e d i c t i o n s f o r t r a i n i n g d a t a u s i n g t r a i n i n g m o d e l
yhat_modA_train <- predict(modA_train, newdata = train)
yhat_modB_train <- predict(modB_train, newdata = train)
yhat_modC_train <- predict(modC_train, newdata = train)
yhat_modD_train <- predict(modD_train, newdata = train)

# C a l c u l a t e R M S E f o r t e s t i n g d a t a
modA_test_RMSE <- sqrt(sum((y_test - yhat_modA_test)^2) / nrow(test))
modB_test_RMSE <- sqrt(sum((y_test - yhat_modB_test)^2) / nrow(test))
modC_test_RMSE <- sqrt(sum((y_test - yhat_modC_test)^2) / nrow(test))
modD_test_RMSE <- sqrt(sum((y_test - yhat_modD_test)^2) / nrow(test))

modA_train_RMSE <- sqrt(sum((y_train - yhat_modA_train)^2) / nrow(train))
modB_train_RMSE <- sqrt(sum((y_train - yhat_modB_train)^2) / nrow(train))
modC_train_RMSE <- sqrt(sum((y_train - yhat_modC_train)^2) / nrow(train))
modD_train_RMSE <- sqrt(sum((y_train - yhat_modD_train)^2) / nrow(train))

data_frame(Model = c("A","B","C","D"),
        RMSE_testdata = c(modA_test_RMSE, modB_test_RMSE,
                modC_test_RMSE, modD_test_RMSE),
        RMSE_traindata = c(modA_train_RMSE, modB_train_RMSE,
                 modC_train_RMSE, modD_train_RMSE),
        ratio_of_RMSEs = RMSE_traindata / RMSE_testdata)

## # A tibble: 4 x 4
##   Model RMSE_testdata RMSE_traindata ratio_of_RMSEs
##   <chr>       <dbl>        <dbl>        <dbl>
## 1 A            1.89         1.86        0.981
## 2 B            3.25         2.98        0.916
## 3 C            1.75         1.68        0.958
## 4 D            1.79         1.60        0.894
                ->most usefulk
Which model makes the most accurate predictions, on average?
Look at RMSE_testdata -> model C (D is close though)
Is there evidence of overfitting?
Not too much, but maybe a bit in model D
```