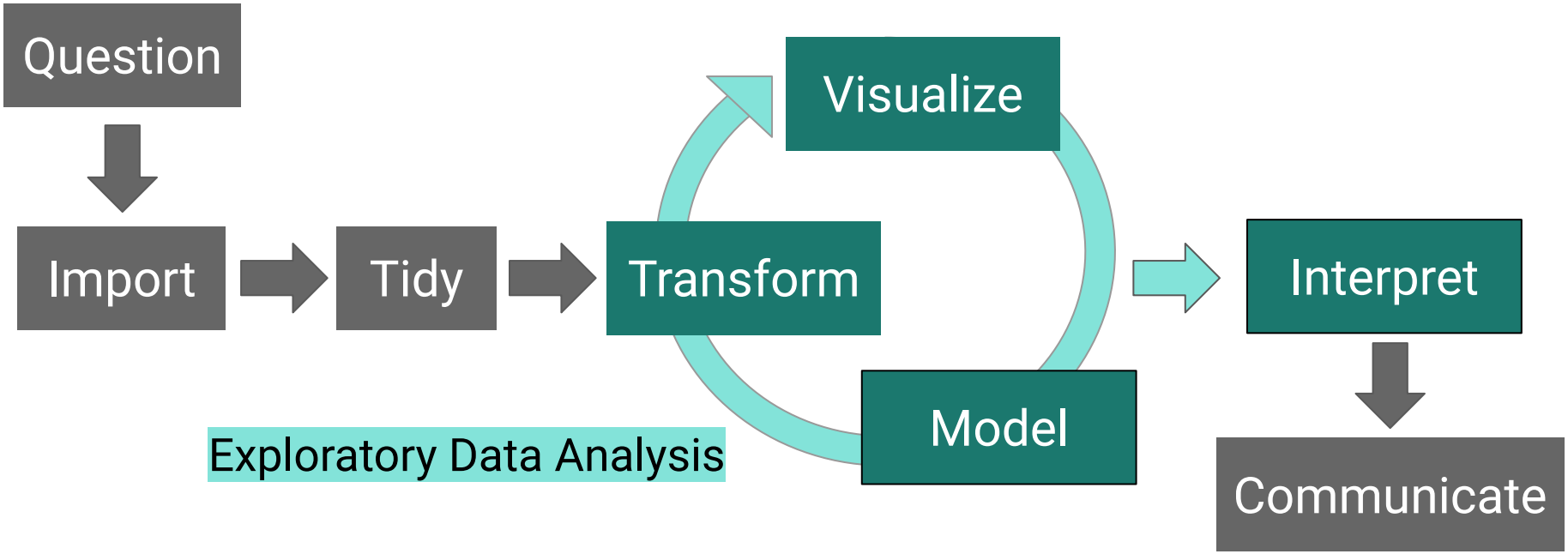
Statistical Modeling II

Lecture 12

Motivation



Modified from https://r4ds.had.co.nz/explore-intro.html#explore-intro

Tidymodels

Tidymodels PACKAGES GET STARTED LEARN HELP CONTRIBUTE

TIDYMODELS

The tidymodels framework is a collection of packages for modeling and machine learning using **tidyverse** principles.

Install tidymodels with:

```
install.packages("tidymodels")
```

Multiple Linear Regression Models

- To model a numerical response or outcome variable using more than one explanatory variable

Multiple Linear Regression Models

- $y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$
- *response* ~ *explanatory variables*
- *response* ~ $variable_1 + variable_2 + \ldots + variable_n$

Babies dataset in R

- The Child Health and Development Studies (USA) investigated the pregnancy in women in San Francisco, 1960-1967. Studied the relationship between mothers who smoked and weight of their babies.
- “babies” dataset available in “openintro” package in R
- 1,236 observations (rows) x 8 variables (columns)

variable	description
case	id number
bwt	birthweight, ounces
gestation	length of gestation, days
parity	binary indicator for a first pregnancy (0=first pregnancy)
age	mother’s age, years
height	mother’s height, inches
weight	mother’s weight, pounds
smoke	binary indicator whether the mother smoked, 1=smoker

Multivariable regression model on birthweight

- Using base R

```
babies %>%
    lm(formula = bwt ~ smoke + gestation + parity +
        age + height + weight)
```

Multivariable regression model on birthweight

- Using tidymodels
  - Step 1: Specify the model
  - Step 2: Run the model
  - Step 3: Analyze the results

Multivariable regression model on birthweight

- Using tidymodels
  - Step 1: Specify the model
  - **linear_reg( )** function is equivalent to **lm( )**
  - **set_engine( )** function is used to specify which package or system will be used to fit the model (e.g. linear regression, random forest, etc.)

```
lm_model <- linear_reg(mode = "regression") %>%
                    set_engine("lm")
```

Multivariable regression model on birthweight

- Using tidymodels
  - Step 2: Run the model using **fit( )** function

```
# Model formula
formula <- bwt ~ smoke + gestation + parity + age + height
+ weight

# Run the model
lm_fit <- lm_model %>%
                fit(formula, data = babies)
```

Multivariable regression model on birthweight

- Using tidymodels
  - Step 3: Access the results using the **pluck( )** and **summary( )**

```
lm_model_res <- lm_fit %>%
                pluck("fit") %>%
                summary()
lm_model_res
```

Multivariable regression model on birthweight

```
Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept) -80.41085    14.34657  -5.605 2.60e-08 ***
smoke        -8.40073     0.95382  -8.807  < 2e-16 ***
gestation     0.44398     0.02910  15.258  < 2e-16 ***
parity       -3.32720     1.12895  -2.947  0.00327 **
age          -0.00895     0.08582  -0.104  0.91696
height        1.15402     0.20502   5.629 2.27e-08 ***
weight        0.05017     0.02524   1.987  0.04711 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.83 on 1167 degrees of freedom
  (62 observations deleted due to missingness)
Multiple R-squared:  0.258,    Adjusted R-squared:  0.2541
F-statistic: 67.61 on 6 and 1167 DF,  p-value: < 2.2e-16
```

How to interpret these results?

Multivariable regression model on birthweight

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-80.41085	14.34657	-5.605	2.60e-08	***
smoke	-8.40073	0.95382	-8.807	< 2e-16	***
gestation	0.44398	0.02910	15.258	< 2e-16	***
parity	-3.32720	1.12895	-2.947	0.00327	**
age	-0.00895	0.08582	-0.104	0.91696	
height	1.15402	0.20502	5.629	2.27e-08	***
weight	0.05017	0.02524	1.987	0.04711	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.83 on 1167 degrees of freedom
(62 observations deleted due to missingness)
Multiple R-squared: 0.258, Adjusted R-squared: 0.2541
F-statistic: 67.61 on 6 and 1167 DF, p-value: < 2.2e-16

β_{smoke}: "The birth weight of newborns for mothers who were smokers is associated with an 8.4 oz decrease on average in weight than nonsmokers, after adjusting for other variables."

Multivariable regression model on birthweight

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-80.41085	14.34657	-5.605	2.60e-08	***
smoke	-8.40073	0.95382	-8.807	< 2e-16	***
gestation	0.44398	0.02910	15.258	< 2e-16	***
parity	-3.32720	1.12895	-2.947	0.00327	**
age	-0.00895	0.08582	-0.104	0.91696	
height	1.15402	0.20502	5.629	2.27e-08	***
weight	0.05017	0.02524	1.987	0.04711	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.83 on 1167 degrees of freedom
(62 observations deleted due to missingness)
Multiple R-squared: 0.258, Adjusted R-squared: 0.2541
F-statistic: 67.61 on 6 and 1167 DF, p-value: < 2.2e-16

$\beta_{gestation}$: "For each additional day of pregnancy is associated with a 0.44 oz increase on average in birth weight of newborns, holding all other variables constant."

Adjusted R^2 as estimate of explained variance

```
Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept) -80.41085    14.34657  -5.605 2.60e-08 ***
smoke        -8.40073     0.95382  -8.807  < 2e-16 ***
gestation     0.44398     0.02910  15.258  < 2e-16 ***
parity       -3.32720     1.12895  -2.947  0.00327 **
age          -0.00895     0.08582  -0.104  0.91696
height        1.15402     0.20502   5.629 2.27e-08 ***
weight        0.05017     0.02524   1.987  0.04711 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.83 on 1167 degrees of freedom
  (62 observations deleted due to missingness)
Multiple R-squared:  0.258,    Adjusted R-squared:  0.2541
F-statistic: 67.61 on 6 and 1167 DF,  p-value: < 2.2e-16
```

“About 25.4% of the variance in baby weight is explained by the data.”

Calculating the 95% confidence interval

- Confidence intervals for coefficients in multiple regression can be computed as:

$$\textbf{95\% confidence interval} = b \pm t_{df} * SE$$

- where **b** is coefficient; t_{df} is the t-value corresponding to the confidence level and model degrees of freedom (df); and **SE** is the standard error
- for example: 95% CI for smoke

$$95\%CI = -8.40073 \pm (-8.807)(0.95382) = [-16.8, 0.0]$$

"We are 95% confident that birth weight of babies from mothers who smoked on average are 16.8 oz. lighter to 0 oz. heavier than nonsmokers when controlling for other variables."

Model selection

- The best model is not always the most complicated.
- Sometimes including variables that are not evidently important can actually reduce the accuracy of predictions.
- In practice, the model that includes all available explanatory variables is often referred to as the **full model**.
- **Adjusted R^2** describes the strength of a model fit.
- Useful tool for evaluating which predictors are adding value to the model.

Backward elimination

1. Start with a full model
2. Drop one variable at a time and a record adjusted R^2 of each smaller model
3. Pick the model with the highest increase in adjusted R^2
4. Repeat until none of the models yield an increase in adjusted R^2

Forward selection

1. Start with regressions of response vs. each explanatory variable
2. Pick the model with the highest adjusted R^2 value
3. Add variables to the existing model one at a time, and pick the model with the highest adjusted R^2 value
4. Repeat until the addition of any remaining variables does not result in a higher adjusted R^2 value

P-value approach

Backward elimination with the p-value approach:

1. Start with the full model
2. Drop the variable with the highest p-value and refit a smaller model
3. Repeat until all variables left in the model are significant

Forward selection with the p-value approach:

1. Start with regressions of response vs. each explanatory variable
2. Pick the variable with the lowest significant p-value
3. Add the remaining variables one at a time to the existing model, and pick the variable with the lowest significant p-value
4. Repeat until any of the remaining variables does not have a significant p-value

step( ) function

- The step function in R does a similar backward elimination process, however it uses a different metric called AIC (Akaike Information Criterion) instead of adjusted R^2 to do the model selection.
- Lowest AIC value is the best-fit model

```
lm1 <- lm(bwt ~ smoke + gestation + parity + age +
                    height + weight, data = babies)
slm1 <- stats::step(lm1, direction = "backward")
```

step( ) function

Start: AIC=6491.82
bwt ~ smoke + gestation + parity + age + height + weight

	Df	Sum of Sq	RSS	AIC
- age	1	3	292412	6489.8
<none>			292409	6491.8
- weight	1	990	293399	6493.8
- parity	1	2176	294586	6498.5
- height	1	7939	300348	6521.3
- smoke	1	19437	311846	6565.4
- gestation	1	58334	350744	6703.4

Step: AIC=6489.83
bwt ~ smoke + gestation + parity + height + weight

	Df	Sum of Sq	RSS	AIC
<none>			292412	6489.8
- weight	1	992	293404	6491.8
- parity	1	2396	294808	6497.4
- height	1	7968	300380	6519.4
- smoke	1	19497	311909	6563.6
- gestation	1	58421	350833	6701.7

step( ) function

summary(slm1)

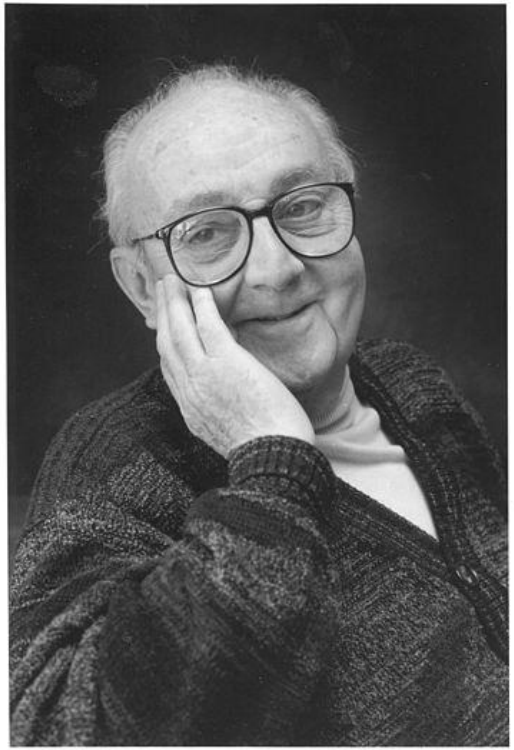
```
Call:
lm(formula = bwt ~ smoke + gestation + parity + height + weight,
    data = babies)

Residuals:
Min      1Q  Median      3Q     Max
-57.716 -10.150  -0.159   9.689  51.620

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept) -80.71321   14.04465  -5.747 1.16e-08 ***
smoke        -8.39390    0.95117  -8.825  < 2e-16 ***
gestation     0.44408    0.02907  15.276  < 2e-16 ***
parity       -3.28762    1.06281  -3.093  0.00203 **
height        1.15497    0.20473   5.641 2.11e-08 ***
weight        0.04983    0.02503   1.991  0.04672 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.82 on 1168 degrees of freedom
Multiple R-squared:  0.2579,    Adjusted R-squared:  0.2548
F-statistic:  81.2 on 5 and 1168 DF,  p-value: < 2.2e-16
```

On modeling

- “All models are wrong, but some are useful.”

  - **George Box**

- No model is perfect, but even imperfect models can be useful, as long as it is clear and report the model’s shortcomings.



https://en.wikipedia.org/wiki/George_E._P._Box

Takeaway message

- Use multiple regression analysis to assess the relationship between a response or outcome variable and several explanatory variables simultaneously
- “All models are wrong, *but some are pure garbage*”