Course Summary

Fundamental Techniques in Data Science

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

R Topics

R Fundamentals

Objects and assignment

```
1:3
```

```
[1] 1 2 3
```

```
x <- 1:3
x
```

```
[1] 1 2 3
```

```
x + 4
```

```
[1] 5 6 7
```

Data types

- Vectors, Matrices
- Lists, Data frames
- Factors

R Fundamentals

User-defined functions

```
helloWorld <- function() cat("Hello World!")
helloWorld()

Hello World!

add <- function(x, y) x + y
add(2, 3)

[1] 5

add(add(1, 2), 3)

[1] 6
```

Tidyverse Fundamentals

Working with pipes

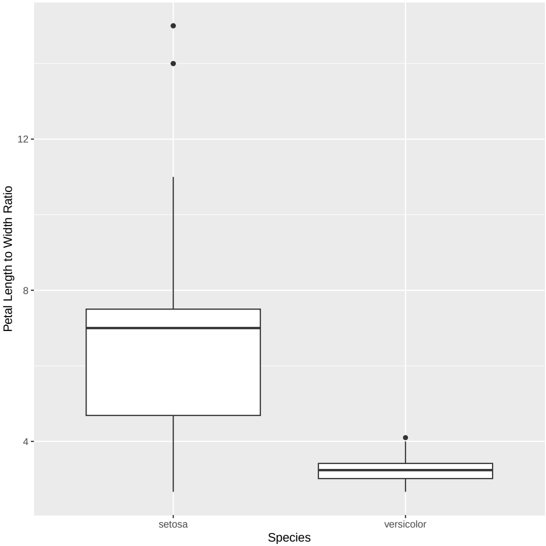
```
library(magrittr)

iris %$% table(Species)

Species
    setosa versicolor  virginica
        50         50         50

add(1, 2) %>% add(3)

[1] 6
```

Tidyverse Fundamentals

Working with **dplyr** and **ggplot**

```
library(dplyr)
library(ggplot2)

iris %>%
    filter(Species != "virginica") %>%
    mutate(petal_ratio = Petal.Length / Petal.Width) %>%
    ggplot(aes(Species, petal_ratio)) +
    geom_boxplot() +
    ylab("Petal Length to Width Ratio")
```

Tidyverse Fundamentals



Manipulating Model Objects

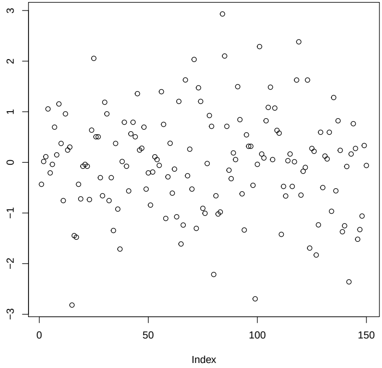
```r
fit1 <- lm(Petal.Length ~ Sepal.Length + Species, data = iris)
fit2 <- lm(Petal.Length ~ Sepal.Length*Species, data = iris)

coef(fit1)

      (Intercept)      Sepal.Length Speciesversicolor  Speciesvirginica
       -1.7023422         0.6321099         2.2101378         3.0900021

summary(fit2)$fstatistic

   value    numdf    dendf
1333.265    5.000  144.000
```

Manipulating Model Objects

```
anova(fit2, fit1)

Analysis of Variance Table

Model 1: Petal.Length ~ Sepal.Length * Species
Model 2: Petal.Length ~ Sepal.Length + Species
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    144  9.8179
2    146 11.6571 -2   -1.8393 13.489 4.272e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Manipulating Model Objects

```
fit1 %>% rstudent() %>% plot()
```



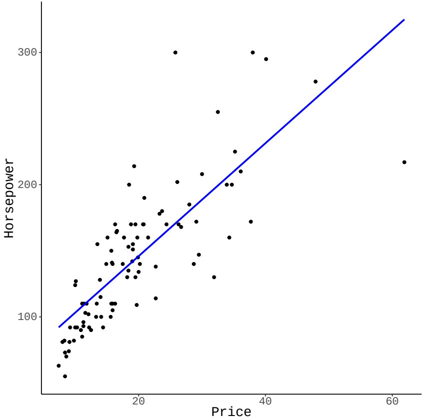
LINEAR REGRESSION

Simple Linear Regression

In linear regression, we want to find the best fit line:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- For any X_n, the corresponding $\hat{Y}_n$ represents the model-implied, conditional mean of Y.



Simple Linear Regression

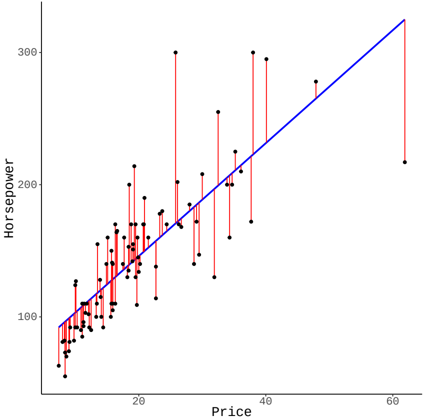
In linear regression, we want to find the best fit line:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- For any X_n, the corresponding $\hat{Y}_n$ represents the model-implied, conditional mean of Y.

After accounting for the estimation error, we get the full regression equation:

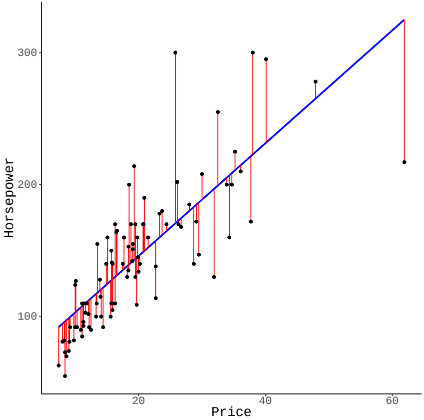
$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\varepsilon}$$



Residuals as the Basis of Estimation

We use the residuals, $\hat{\varepsilon}_n$, to estimate the model.

$$RSS = \sum_{n=1}^{N} \hat{\varepsilon}_n^2 = \sum_{n=1}^{N} \left(Y_n - \hat{Y}_n\right)^2$$
$$= \sum_{n=1}^{N} \left(Y_n - \hat{\beta}_0 - \hat{\beta}_1 X_n\right)^2$$



Sampling Distributions & Standard Errors

Fit a simple linear regression model to some toy data.

```r
library(mvtnorm)
library(magrittr)

set.seed(235711)

s <- matrix(c(1.0, 0.5,
              0.5, 1.0),
            ncol = 2)

fit <- rmvnorm(n = 1000, mean = c(0, 0), sigma = s) %>%
    data.frame() %$%
    lm(X1 ~ X2)
```

Sampling Distributions & Standard Errors

```
summary(fit)


Call:
lm(formula = X1 ~ X2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.85250 -0.58340  0.00822  0.58493  2.56188 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.03099    0.02728   1.136    0.256    
X2           0.43143    0.02761  15.626   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8626 on 998 degrees of freedom
Multiple R-squared:  0.1966,Adjusted R-squared:  0.1958 
F-statistic: 244.2 on 1 and 998 DF,  p-value: < 2.2e-16
```

Sampling Distributions & Standard Errors

Calculate the residual variance.

```
## Extract residuals:
res <- resid(fit)

## Compute the residual sum of squares:
ssr <- {res - mean(res)}^2 %>% sum()

## Compute the residual variance:
ssr / fit$df.residual

[1] 0.7440214

summary(fit)$sigma^2

[1] 0.7440214
```

Sampling Distributions & Standard Errors

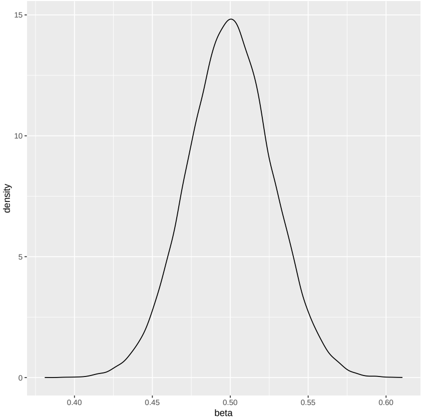
Repeat the above analysis many times.

```
set.seed(235711)

nReps <- 50000
betas <- rep(NA, nReps)

for(i in 1:nReps)
    betas[i] <- rmvnorm(n = 1000, mean = c(0, 0), sigma = s) %>%
        data.frame() %$%
        lm(X1 ~ X2) %>%
        coef() %>%
        purrr::pluck(2)
```

Sampling Distributions & Standard Errors

Visualize the slope's sampling distribution.

```
ggplot(data.frame(beta = betas), aes(beta)) + geom_density()
```



Sampling Distributions & Standard Errors

Estimate the slope's standard error.

```
## The Monte Carlo approximation:
sd(betas)

[1] 0.02735199

## The asymptotic estimate:
summary(fit)$coefficients[2, "Std. Error"]

[1] 0.02761019
```

Example

```
## Read in the 'diabetes' dataset:
diabetes <- readRDS("../data/diabetes.rds")

## Estimate and summarize a regression model:
lm(bp ~ age + ldl + hdl + sex, data = diabetes) %>% partSummary(-1)

Residuals:
    Min      1Q  Median      3Q     Max
-34.195  -8.734  -1.011   7.945  42.186

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 78.18713    4.29453  18.206  < 2e-16
age          0.30043    0.04789   6.273 8.52e-10
ldl          0.03887    0.02079   1.870  0.06220
hdl         -0.09063    0.05124  -1.769  0.07763
sexmale      4.07606    1.32803   3.069  0.00228

Residual standard error: 12.72 on 437 degrees of freedom
Multiple R-squared:  0.162,Adjusted R-squared:  0.1543
F-statistic: 21.12 on 4 and 437 DF,  p-value: 6.163e-16
```

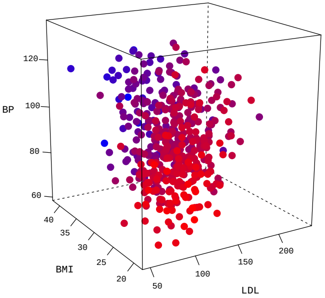
Assumptions

1. The model is linear in the parameters.
   - *Otherwise:* We are not working with linear regression.
2. The predictor matrix is *full rank*.
   - *Otherwise:* The model is not estimable.
3. The predictors are strictly exogenous.
   - *Otherwise:* The estimated regression coefficients will be biased.
4. The errors have constant, finite variance.
   - *Otherwise:* Standard errors will be biased.
5. The errors are uncorrelated.
   - *Otherwise:* Standard errors will be biased.
6. The errors are normally distributed.
   - *Otherwise:* Small-sample inferences and some estimates are not justified.

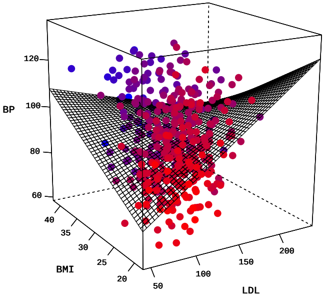
MODERATION

Moderated Regression

The effect of X on Y varies **as a function** of Z.



$\longrightarrow$



Interpretation

Given the following equation:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 Z + \hat{\beta}_3 XZ + \hat{\varepsilon}$$

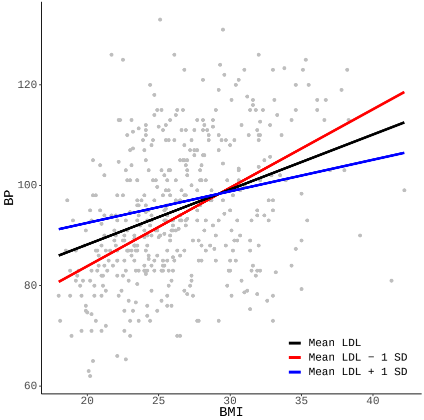
- $\hat{\beta}_3$ quantifies the effect of Z on the focal effect (the $X \rightarrow Y$ effect).
  - For a unit change in Z, $\hat{\beta}_3$ is the expected change in the effect of X on Y.
- $\hat{\beta}_1$ and $\hat{\beta}_2$ are *conditional effects*.
  - Interpreted where the other predictor is zero.
  - For a unit change in X, $\hat{\beta}_1$ is the expected change in Y, when $Z = 0$.
  - For a unit change in Z, $\hat{\beta}_2$ is the expected change in Y, when $X = 0$.

Continuous Moderators

```
## Moderated Model:
out2 <- lm(bp ~ bmi * ldl, data = diabetes)
partSummary(out2, -c(1, 2))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.480616  14.291677   1.013 0.311514
bmi          2.867825   0.541312   5.298 1.86e-07
ldl          0.448771   0.127160   3.529 0.000461
bmi:ldl     -0.015352   0.004716  -3.255 0.001221

Residual standard error: 12.54 on 438 degrees of freedom
Multiple R-squared:  0.1834,Adjusted R-squared:  0.1778
F-statistic: 32.78 on 3 and 438 DF,  p-value: < 2.2e-16
```

Visualizing the Interaction

We can get a better idea of the patterns of moderation by plotting the focal effect at conditional values of the moderator.



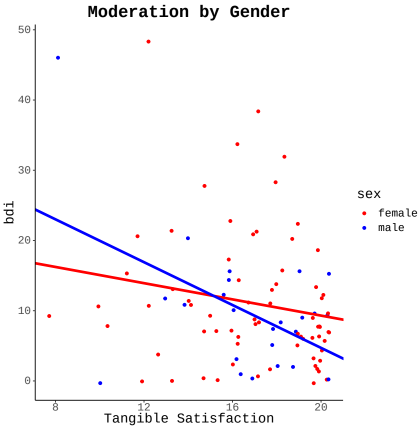
Categorical Moderators

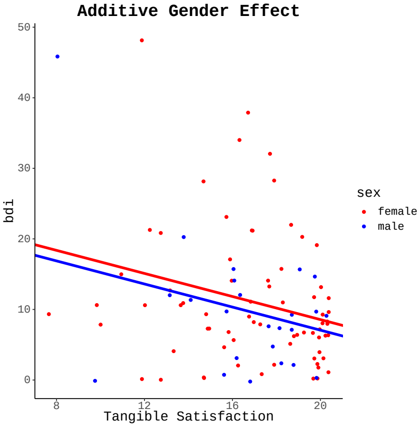
```
## Load data:
socSup <- readRDS("../data/social_support.rds")

## Estimate the moderated regression model:
out4 <- lm(bdi ~ tanSat * sex, data = socSup)
partSummary(out4, -c(1, 2))

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     20.8478     6.2114   3.356  0.00115
tanSat          -0.5772     0.3614  -1.597  0.11372
sexmale         14.3667    12.2054   1.177  0.24223
tanSat:sexmale  -0.9482     0.7177  -1.321  0.18978

Residual standard error: 9.267 on 91 degrees of freedom
Multiple R-squared:  0.08955,Adjusted R-squared:  0.05954
F-statistic: 2.984 on 3 and 91 DF,  p-value: 0.03537
```

Visualizing Categorical Moderation

$$\hat{Y}_{BDI} = 20.85 - 0.58X_{tsat} + 14.37Z_{male} - 0.95X_{tsat}Z_{male}$$

$$\hat{Y}_{BDI} = 24.91 - 0.82X_{tsat} - 1.50Z_{male}$$





PREDICTION

Prediction Example

Let's fit the following model using the *diabetes* data:

$$Y_{LDL} = \beta_0 + \beta_1 X_{BP} + \beta_2 X_{gluc} + \beta_3 X_{BMI} + \varepsilon$$

Training this model on the first $N = 400$ patients' data produces the following fitted model:

$$\hat{Y}_{LDL} = 22.135 + 0.089 X_{BP} + 0.498 X_{gluc} + 1.48 X_{BMI}$$

Prediction Example

Let's fit the following model using the *diabetes* data:

$$Y_{LDL} = \beta_0 + \beta_1 X_{BP} + \beta_2 X_{gluc} + \beta_3 X_{BMI} + \varepsilon$$

Training this model on the first $N = 400$ patients' data produces the following fitted model:

$$\hat{Y}_{LDL} = 22.135 + 0.089 X_{BP} + 0.498 X_{gluc} + 1.48 X_{BMI}$$

Suppose a new patient presents with $BP = 121$, $gluc = 89$, and $BMI = 30.6$. We can predict their LDL score by:

$$\begin{aligned}\hat{Y}_{LDL} &= 22.135 + 0.089(121) + 0.498(89) + 1.48(30.6) \\ &= 122.463\end{aligned}$$

Interval Estimates Example

Two flavors of interval to quantify prediction uncertainty:

1. Confidence intervals
2. Prediction intervals

In our example, we get the following 95% interval estimates:

$$95\%\ CI_{\hat{Y}} = [115.6; 129.33]$$

$$95\%\ PI = [66.56; 178.37]$$

- We can be 95% confident that the average *LDL* of patients with *Glucose* = 89, *BP* = 121, and *BMI* = 30.6 will be somewhere between 115.6 and 129.33.

- We can be 95% confident that the *LDL* of a specific patient with *Glucose* = 89, *BP* = 121, and *BMI* = 30.6 will be somewhere between 66.56 and 178.37.

MODEL FIT

Model Fit

We quantify the proportion of the outcome's variance that is explained by our model using the R^2 statistic:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

where

$$TSS = \sum_{n=1}^{N} \left(Y_n - \bar{Y}\right)^2 = \mathrm{Var}(Y) \times (N - 1)$$

For the model we estimated in the above prediction example, we get:

$$R^2 = 1 - \frac{315383}{361704} \approx 0.13$$

Model Fit for Prediction

We use the *mean squared error* (MSE) to assess predictive performance.

$$
\begin{aligned}
MSE &= \frac{1}{N} \sum_{n=1}^{N} \left(Y_n - \hat{Y}_n \right)^2 \\
&= \frac{1}{N} \sum_{n=1}^{N} \left(Y_n - \hat{\beta}_0 - \sum_{p=1}^{P} \hat{\beta}_p X_{np} \right)^2 \\
&= \frac{RSS}{N}
\end{aligned}
$$

For our example problem, we get:

$$
MSE = \frac{315383}{400} \approx 788.46
$$

Information Criteria

We can use *information criteria* to quickly compare *non-nested* (or nested) models while accounting for model complexity.

- Akaike's Information Criterion (AIC)

$$AIC = 2K - 2\hat{\ell}(\theta|X)$$

- Bayesian Information Criterion (BIC)

$$BIC = K\ln(N) - 2\hat{\ell}(\theta|X)$$

For our example, we get the following estimates of AIC and BIC:

$$\begin{aligned} AIC &= 2(3) - 2(-1901.59) \\ &= 3813.18 \end{aligned}$$

$$\begin{aligned} BIC &= 3\ln(400) - 2(-1901.59) \\ &= 3833.14 \end{aligned}$$

Logistic Regression

Probabilities & Odds

	Complete	
Sex	No	Yes
Female	95	147
Male	753	1540

$$P(C|M) = \frac{1540}{1540 + 753} = 0.672$$

$$O(C|M) = \frac{1540}{753} = 2.045 \approx \frac{0.672}{1 - 0.672}$$

$$P(C|F) = \frac{147}{147 + 95} = 0.607$$

$$O(C|F) = \frac{147}{95} = 1.547 \approx \frac{0.607}{1 - 0.607}$$

The Generalized Linear Model

Every GLM is built from three components:

1. The systematic component, η.
    - A linear function of the predictors, $\{X_p\}$.
    - Describes the association between $\mathbf{X}$ and Y.

2. The link function, $g(\mu_Y)$.
    - Transforms μ_Y so that it can take any value on the real line.

3. The random component, $P(Y|g^{-1}(\eta))$
    - The distribution of the observed Y.
    - Quantifies the error variance around η.

The Logistic Regression Model

The logistic regression model can be represented as:

$$Y \sim \mathrm{Bin}(\pi, 1)$$

$$\mathrm{logit}(\pi) = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$$

The fitted model can be represented as:

$$\mathrm{logit}(\hat{\pi}) = \hat{\beta}_0 + \sum_{p=1}^{P} \hat{\beta}_p X_p$$

To convert fitted values, $\hat{\eta} = \hat{\beta}_0 + \sum_{p=1}^{P} \hat{\beta}_p X_p$, from a logit scale to a probability scale, we apply the *logistic* function:

$$\mathrm{logistic}(\hat{\eta}) = \frac{e^{\hat{\eta}}}{1 + e^{\hat{\eta}}}$$

Logistic Regression Example

```
## Coarsen the blood glucose variable:
diabetes %<>% mutate(highGlu = as.numeric(glu > 90))

## Estimate the model:
out1 <- glm(highGlu ~ age + bmi + bp, data = diabetes, family = binomial())
partSummary(out1, -c(1, 2))

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.479104   0.912899  -7.097 1.27e-12
age          0.034597   0.008635   4.007 6.16e-05
bmi          0.106852   0.026660   4.008 6.12e-05
bp           0.022691   0.008560   2.651  0.00803

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 610.42  on 441  degrees of freedom
Residual deviance: 538.18  on 438  degrees of freedom
AIC: 546.18

Number of Fisher Scoring iterations: 4
```

Assumptions

We can state the assumptions of logistic regression as follows:

1. The outcome follows a binomial distribution.
2. The predictor matrix is full-rank.
3. The predictors are linearly related to $logit(\pi)$.
4. The observations are independent after accounting for the predictors.

Unlike linear regression, we don't need to assume

- Constant, finite error variance
- Normally distributed errors

For computational reasons, we also need the following:

- Large sample
- Relatively well-balance outcome
- No highly influential cases

CLASSIFICATION

Classification Example

Say we want to classify a new patient into either the “high glucose” group or the “not high glucose” group using the model fit above.

- Assume this patient has the following characteristics:
  - They are 57 years old
  - Their BMI is 28
  - Their average blood pressure is 92

First we plug their predictor data into the fitted model to get their model-implied η:

$$\begin{aligned}\hat{\eta} &= -6.479 + 0.035 \times 57 + 0.107 \times 28 + 0.023 \times 92 \\ &= 0.572\end{aligned}$$

Classification Example

Next we convert the predicted η value into a model-implied success probability by applying the logistic function:

$$\hat{\pi} = \text{logistic}(0.572) = \frac{e^{0.572}}{1 + e^{0.572}} = 0.639$$

Finally, to make the classification, assume a threshold of $\hat{\pi} = 0.5$ as the decision boundary.

- Because $0.639 > 0.5$ we would classify this patient into the “high glucose” group.

Confusion Matrix

	Predicted	
True	Low	High
Low	123	82
High	62	175

Confusion Matrix of Blood Glucose Level

$$Sensitivity = \frac{175}{175 + 62} = 0.738$$

$$Specificity = \frac{123}{123 + 82} = 0.6$$

$$Accuracy = \frac{175 + 123}{175 + 123 + 62 + 82} = 0.674$$

ROC Curve