NUS National University of Singapore | School of Computing
Leading The World With Asia's Best

Introduction to Business Analytics

Linear Regression (2)

Dr. Sharon Tan

Concept review

- For this simple linear regression model:

  $SaleRevenue = b_0 + b_1$*MktingExpenditure

  - Both SaleRevenue and MktingExpenditure are in $/year.

  - We find $b_0 = -2500$ and $b_1 = 25.8$. How do we interpret this?

Concept review

- For this multiple linear regression model:

  SaleRevenue = b_0 + b_1*MktingExpenditure + b_2*CustService

  - Both SaleRevenue and MktingExpenditure are in $/year and CustService is the based on survey rating for that year (0-10 with higher rating being better customer service).
  - We find $b_0 = -1800$, $b_1 = 18.9$ and $b_2 = 6578$. What do these tell us?
  - If a company was to spend $10000/year on marketing and scores a 10 for the customer service survey, what is the average sales revenue that the company is predicted to obtain?

Concept review

- For this simple linear regression model:

  $SalesRevenue = b_0 + b_1 * PriceStrategy$

  - SaleRevenue is in $/year customer spends while PriceStrategy is a categorical variable with PriceStrategy =1 when the company uses pricing strategy and 0 if the company does not use any pricing strategy.

  - We find $b_0 = 6500$ and $b_1 = 3698$. What do these mean?

Concept review

- For this logistic regression model, we want to predict whether a company makes a profit (yes/no):

$$\text{logit(Profit)} = b_0 + b_1\text{*PriceStrategy} + b_2\text{*CustService}$$

- CustService is the based on survey rating for that year (0-10 with higher rating being better customer service).

- PriceStrategy is a categorical variable with PriceStrategy =1 when the company uses pricing strategy and 0 if the company does not use any pricing strategy.

- We find $b_0 = -2$, $b_1 = 0.5$ and $b_2 = 0.3$.

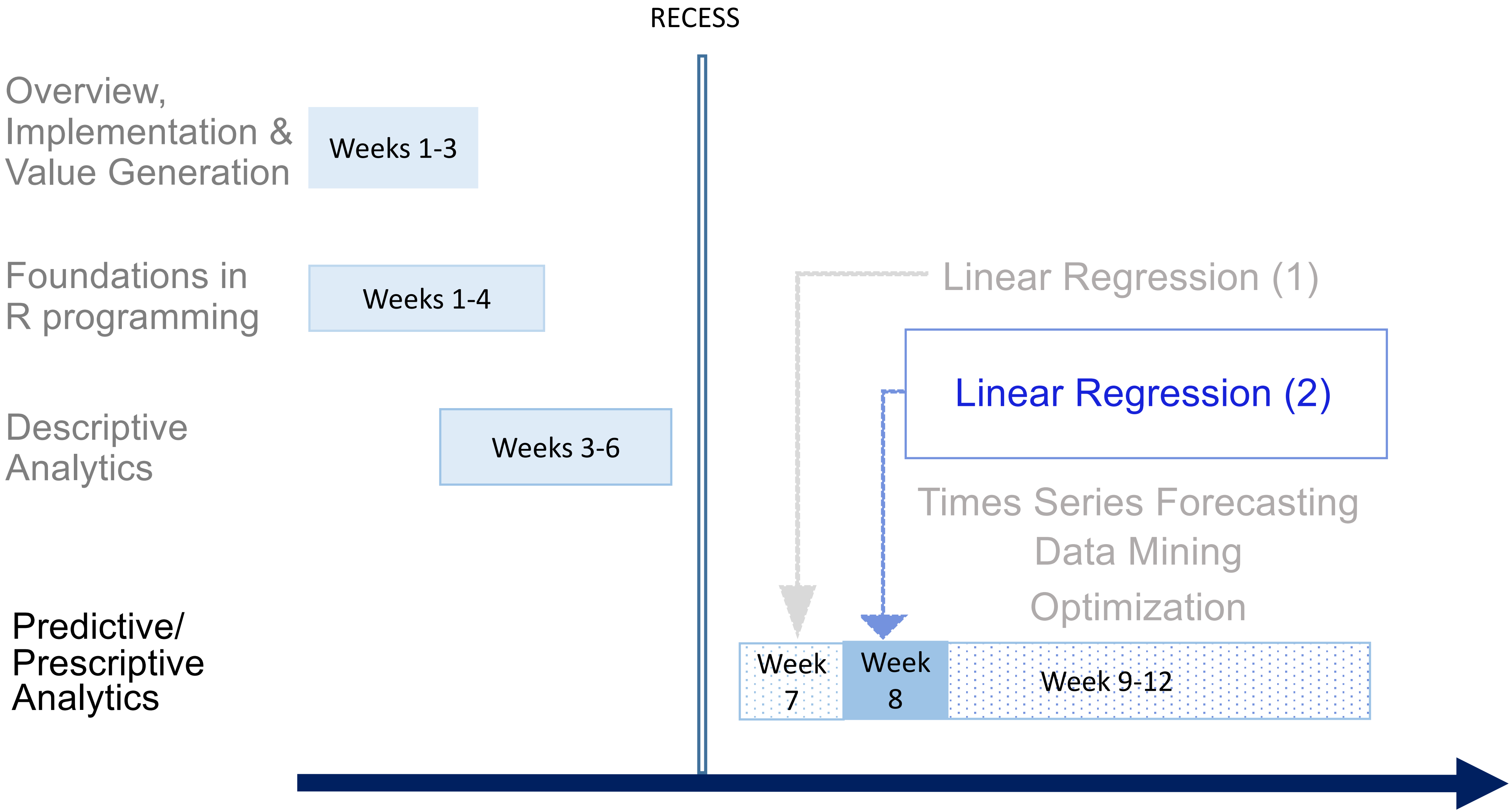
- What do these mean?

Concept review

- how do we interpret these goodness-of-fit statistics?

```
Residual standard error: 28.39 on 151 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.08573,	Adjusted R-squared:  0.07968
F-statistic: 14.16 on 1 and 151 DF,  p-value: 0.0002398
```

Course Map


RECESS
Overview,
Implementation &
Value Generation
Weeks 1-3
Foundations in
R programming
Weeks 1-4
Descriptive
Analytics
Weeks 3-6
Linear Regression (1)
Linear Regression (2)
Times Series Forecasting
Data Mining
Optimization
Predictive/
Prescriptive
Analytics
Week
7
Week
8
Week 9-12

Learning outcomes

- Test for interactions in regression models with categorical variables
- Predict with linear models
- Apply a systematic approach to build good regression models
- Explain the importance of understanding multicollinearity in regression models

Interactions in Multiple Regression

- Interaction occurs when the effect of one variable (i.e. the slope) is dependent on another variable (called the moderator).
- It is modelled by adding an interaction term ($X_1 * X_2$) into the linear regression model
- For example, for this multiple linear regression model (the effects of MktingExpenditure is dependent on PricingStrategy,

$$\text{SaleRevenue} = b_0 + b_1\text{*MktingExpenditure} + b_2\text{*PriceStrategy}$$
$$+ b_3\text{*(MktingExpenditure X PriceStrategy)}$$

- This means:

PriceStrategy = 1: $\text{SaleRevenue} = b_0 + b_2 + (b_1 + b_3)\text{*MktingExpenditure}$

PriceStrategy = 0: $\text{SaleRevenue} = b_0 + b_1\text{*MktingExpenditure}$

- Interpretation:

  b_0 = The average SaleRevenue when there is no pricing strategy and zero marketing expenditure

  b_1 = In the absence of pricing strategy, every \$1/year increase in marketing expenditure, average sales revenue also changes by \$$b_1$ per year. (-b1 means decrease; +b1 means increase)

  b_2 = The average difference in SaleRevenue when there is pricing strategy versus no pricing strategy and when MktingExpenditure is 0.

  b_3= This is the additional change in average sales revenue per year for every \$1/yr increase in marketing expenditure, when there is a pricing strategy versus no pricing strategy.

- Here, we say PriceStrategy is the moderator and if b_3 is statistically different from 0, then we say it has a significant moderation effect on Mkting Expenditure (or the interaction term is significant).

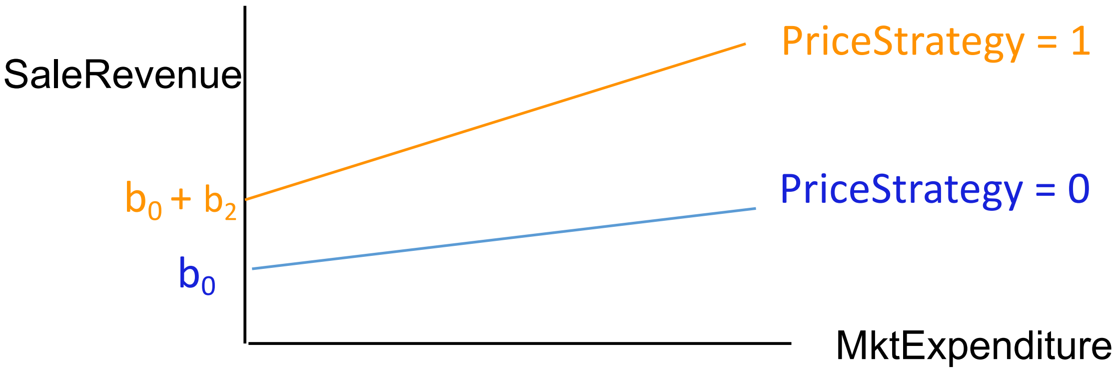
Interactions in Multiple Regression

SaleRevenue = b_0 + b_1*MktingExpenditure + b_2*PriceStrategy + b_3*(MktingExpenditure X PriceStrategy)

- This means:

PriceStrategy = 1: SaleRevenue = $b_0 + b_2 + (b_1 + b_3)$*MktingExpenditure

PriceStrategy = 0: SaleRevenue = $b_0 + b_1$*MktingExpenditure



- In linear regression, we test interaction or moderation by adding the product term into the linear model
- For eg:

```
lm(y ~ x1*x2, df)
# is shorthand and equivalent to
lm(y ~ x1 + x2 + x1*x2, df)
```

Interactions can also be pairs of continuous variables

Predicting with linear models

What is the predicted Ozone level with Solar Radiation of 210 Langley and Wind speed of 11miles per hour?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 77.24604    9.06751   8.519 1.05e-13 ***
X1           0.10035    0.02628   3.819 0.000224 ***
X2          -5.40180    0.67324  -8.024 1.34e-12 ***
```

```
Y<- airquality[,"Ozone"]
X1<- airquality[,"Solar.R"]
X2<- airquality[,"Wind"]
```

- Method 1: write the equation and substitute coefficients

$$Ozone = b_0 + b_1\text{Solar} + b_2\text{Wind}$$
$$= 77.24604 + 0.10035 * 210 + (-5.40180)*11 = 38.899$$

- Method 2a: Use predict() function with manual input

```
> predict(model2,newdata=data.frame(X1=210, X2=11))
      1 
38.8999 
```

- Method 2b: Use predict() function with a dataset

```
> predict(model2,newdata=testset)
```

Pairwise Model Selection

"Full model": $Y \sim X_1 + X_2$

"Restricted model": $Y \sim X_1$

```
m_full <- lm (y ~ x1*x2, df1)
m_restricted <- lm(y~x1, df1)
# model comparison,
anova(m_restricted, m_full)
```

H_0: b_2 (coefficient on X_2) =0

H_1: b_2 (coefficient on X_2) !=0

- anova function with two lm objects conducts a test to see if the explanatory power of the full model is significantly better than the explanatory power of the restricted model i.e., "Is the full model significantly better?"
- to use an anova, restricted model must be a nested model within the full model (ie. it must be a "subset" of the full model)
- Eg: $Y \sim X_1 + X_2$ [restricted: model without interactions]

  $Y \sim X_1 + X_2 + (X_1 * X_2)$ [full: model with interactions]

Pairwise Model Selection

Example 1:

$$Y \sim X_1 + X_2$$

$$\boxed{Y \sim X_1}$$

```
Analysis of Variance Table

Model 1: y ~ x1
Model 2: y ~ x1 + x2
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     19 74.388
2     18 74.225  1   0.16288 0.0395 0.8447
```

F-statistic   *p*-value

The F-statistic is very small, and the *p*-value is very large.
We cannot reject the null hypothesis (that $b_2 = 0$).

The ANOVA suggests that the full model ('model 2') is not significantly better than the restricted model.

-> Pick the simpler model.

Pairwise Model Selection

Example 2:

$$Y \sim X_1 + X_3$$

$$Y \sim X_1$$

```
Model 1: y ~ x1
Model 2: y ~ x1 + x3
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1     19 74.388                                  
2     18 16.381  1    58.007 63.738 2.523e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What is the interpretation in this case? Which model should we go with?

Model Selection

- **Stepwise Regression**: method often used when there is a large set of variables to choose from

Example:

If we have 3 independent variables – X_1, X_2, X_3

There are 2^3 (or 8) possible models:

$Y \sim X_1 + X_2 + X_3$

$Y \sim X_1 + X_2$

$Y \sim X_2 + X_3$

$Y \sim X_1 + X_3$

$Y \sim X_1$

$Y \sim X_2$

$Y \sim X_3$

The number of possible models can grow quickly if you have large k (e.g. 2^{10} is more than a thousand while or 2^{20} is more than 1 million)

Model Selection

- **Backward Stepwise Regression**: start with model with all variables, then at each step, eliminate a predictor until the model does not improve anymore

$$Y \sim X_1 + X_2 + X_3 + X_4$$

$\downarrow$ compare pairwise with all models with 1 variable removed

$$Y \sim X_1 + X_2 + X_3$$

$\downarrow$ compare pairwise with all models with 1 variable removed

$$Y \sim X_1 \quad + X_3 \longrightarrow$$ **Terminate**

- **Forward Stepwise Regression**: starts with small model (just the intercept) then expands one variable at a time, adding the "best" predictor according to some criterion (such as "lowest p-value", "highest adjusted R2", "lowest Mallow's Cp", "lowest AIC")

- **Forward-backward or mixed stepwise regression:** contemplating both adding and removing one variable at each step, and take the best step

Model Selection

- In R, we use the functions in "**olsrr**" package for Stepwise Regression

  Backward Stepwise Regression: ols_step_forward_p(model)

  Forward Stepwise Regression: ols_step_backward_p(model)

  Forward-Backward Regression: ols_step_both_p(model)

  * replace p by AIC if using lowest AIC instead of lowest p-value as criterion

  * AIC (Akaike information criterion) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. It estimates the relative amount of information lost by a given model; less information lost the higher the quality of model.

  * model is the lm() output

- Using mtcars data as example, we consider 4 IVs (disp, hp, wt, qsec) that could affect mpg. Hence our linear regression model will be as follows:

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
```

```
> mtcars
                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
```

Stepwise Regression: lowest p-value

```
> ols_step_forward_p(model)
                              Selection Summary
--------------------------------------------------------------------------
        Variable                  Adj.
Step    Entered     R-Square    R-Square     C(p)        AIC        RMSE
--------------------------------------------------------------------------
   1    wt            0.7528      0.7446    12.4809    166.0294    3.0459
   2    hp            0.8268      0.8148     2.3690    156.6523    2.5934
   3    qsec          0.8348      0.8171     3.0617    157.1426    2.5778
--------------------------------------------------------------------------
```

```
> ols_step_backward_p(model)
                            Elimination Summary
-------------------------------------------------------------------------
        Variable                  Adj.
Step    Removed     R-Square    R-Square     C(p)       AIC        RMSE
-------------------------------------------------------------------------
   1    disp          0.8348      0.8171    3.0617    157.1426    2.5778
-------------------------------------------------------------------------
```

```
> ols_step_both_p(model)
                              Stepwise Selection Summary
-------------------------------------------------------------------------------------
                     Added/                   Adj.
Step    Variable    Removed     R-Square    R-Square     C(p)        AIC        RMSE
-------------------------------------------------------------------------------------
   1       wt       addition       0.753       0.745    12.4810    166.0294    3.0459
   2       hp       addition       0.827       0.815     2.3690    156.6523    2.5934
-------------------------------------------------------------------------------------
```

Stepwise Regression: lowest AIC

```
> ols_step_forward_aic(model)
                          Selection Summary
-------------------------------------------------------------------
Variable        AIC       Sum Sq        RSS       R-Sq    Adj. R-Sq
-------------------------------------------------------------------
wt            166.029    847.725    278.322    0.75283      0.74459
hp            156.652    930.999    195.048    0.82679      0.81484
-------------------------------------------------------------------
```

```
> ols_step_backward_aic(model)
                    Backward Elimination Summary
-------------------------------------------------------------------
Variable        AIC        RSS      Sum Sq       R-Sq    Adj. R-Sq
-------------------------------------------------------------------
Full Model    159.070    185.635    940.412    0.83514      0.81072
disp          157.143    186.059    939.988    0.83477      0.81706
qsec          156.652    195.048    930.999    0.82679      0.81484
-------------------------------------------------------------------
```

```
> ols_step_both_aic(model)
                            Stepwise Summary
--------------------------------------------------------------------------------
Variable      Method        AIC        RSS       Sum Sq       R-Sq    Adj. R-Sq
--------------------------------------------------------------------------------
wt           addition    166.029    278.322    847.725    0.75283      0.74459
hp           addition    156.652    195.048    930.999    0.82679      0.81484
--------------------------------------------------------------------------------
```

Stepwise Regression: best subset

ols_step_best_subset(): subset of predictors that do the best at meeting some well-defined objective criterion, such as having the largest R2 value or the smallest MSE, Mallow's Cp or AIC.

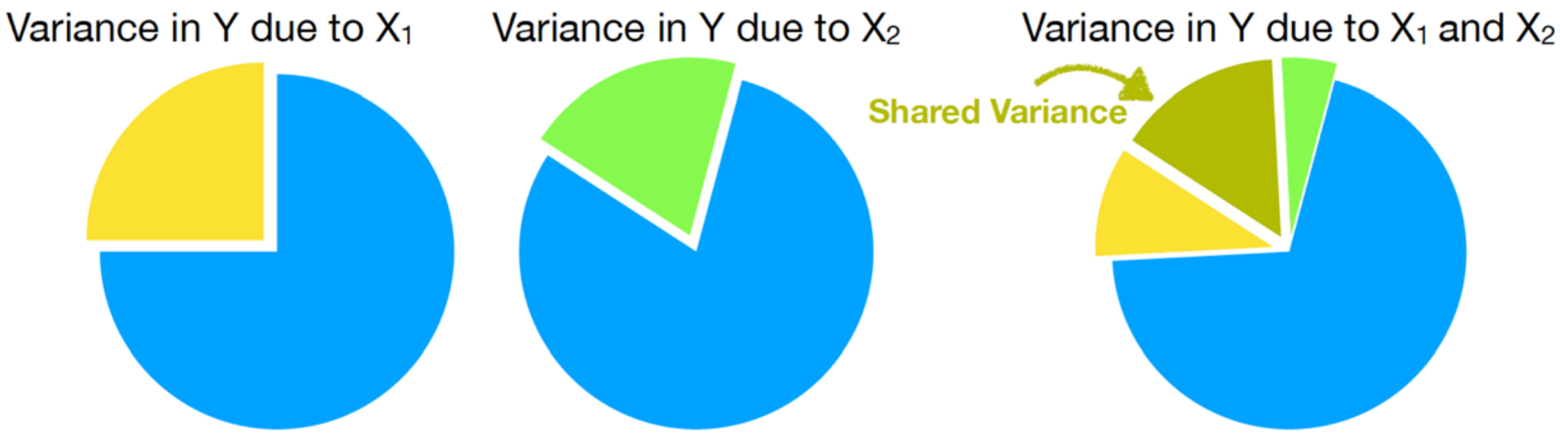
```
> ols_step_best_subset(model)
   Best Subsets Regression
------------------------------
Model Index    Predictors
------------------------------
     1         wt
     2         hp wt
     3         hp wt qsec
     4         disp hp wt qsec
------------------------------

                                                     Subsets Regression Summary
------------------------------------------------------------------------------------------------------------------------------------
                       Adj.        Pred
Model    R-Square    R-Square    R-Square     C(p)        AIC         SBIC        SBC         MSEP        FPE       HSP       APC
------------------------------------------------------------------------------------------------------------------------------------
  1        0.7528      0.7446      0.7087    12.4809    166.0294    74.2916    170.4266    296.9167    9.8572    0.3199    0.2801
  2        0.8268      0.8148      0.7811     2.3690    156.6523    66.5755    162.5153    215.5104    7.3563    0.2402    0.2091
  3        0.8348      0.8171       0.782     3.0617    157.1426    67.7238    164.4713    213.1929    7.4756    0.2461    0.2124
  4        0.8351      0.8107       0.771     5.0000    159.0696    70.0408    167.8640    220.8882    7.9497    0.2644    0.2259
------------------------------------------------------------------------------------------------------------------------------------
AIC: Akaike Information Criteria
 SBIC: Sawa's Bayesian Information Criteria
 SBC: Schwarz Bayesian Criteria
 MSEP: Estimated error of prediction, assuming multivariate normality
 FPE: Final Prediction Error
 HSP: Hocking's Sp
 APC: Amemiya Prediction Criteria
```

* There are many different criteria but no universal agreement on which is the best. Definition/knowledge of these criteria is not within the scope of this module.

Multicollinearity

- Multicollinearity occurs when two or more regressors are "collinear" with each other, that is, they are very highly correlated with each other (e.g. r>0.7 as a rule of thumb)



A large proportion of shared variance means that it is hard to isolate effect of one IV on the DV and to estimate the errors, which leads to inflated estimates of errors (and consequently, unstable models, high errors, and high p-values) that may reject in conclusion not to reject H0 when it should be rejected.

Multicollinearity

In data=dfMC, x4 and x5 are highly correlated (r=0.975)

```
cor(dfMC$x4, dfMC$x5)
[1] 0.9748428
```

Y ~ x4 + x6

```
Call:
lm(formula = y ~ x4 + x6, data = dfMC)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.84844 0.42076 -4.393 0.000351 ***
x4 0.36353 0.03456 10.518 4.08e-09 ***
X6 0.22975 0.24649 0.932 0.363633
```

Y ~ x5 + x6

```
Call:
lm(formula = y ~ x5 + x6, data = dfMC)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.99440 0.24501 8.140 1.91e-07 ***
x5 0.37145 0.03730 9.957 9.54e-09 ***
x6 0.07548 0.26028 0.290 0.775
```

Y ~ x4 + x5 + x6

R2 = 0.869
But none of the predictors are significant

```
lm(formula = y ~ x4 + x5 + x6, data = dfMC)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.5876 1.7217 -0.341 0.737
x4 0.2444 0.1614 1.514 0.148
x5 0.1257 0.1663 0.756 0.460
x6 0.1748 0.2598 0.673 0.510
```

Multicollinearity

- When can multicollinearity happen?
  - When two variables are very similar (e.g., by definition or measurement). E.g., Monthly Household Income and Annual Household Income; Number of customers entering the store, and Number of customers leaving the store; {Number of Customers, Average Spending per Customer} and {Total Customer Spending}; etc.)
  - Variables that sum to a fixed number "Number of people who said yes" and "Number of people who said no"; If I have only 3 groups that sum to 100%: % of Satisfied Customers, % of Neutral Customers, % of Dissatisfied Customers
  - This will also happen when using k dummy variables for a categorical variable with k

Note: if there's perfect multicollinearity (i.e., cor(X_4, X_{4a}) = 1), you should see it in your `lm` model output.

```
Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.81260    0.41755  -4.341 0.000352 ***
x4           0.36667    0.03428  10.697 1.76e-09 ***
x4a               NA         NA      NA       NA
```

Multicollinearity: Applying VIF

- Compute Variance Inflation Factor (VIF) for each IV
  - VIF measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model.
  - The smallest possible value of VIF is one (absence of multicollinearity). As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity (James et al. 2014).

Y ~ x4 + x5 + x6

```
lm(formula = y ~ x4 + x5 + x6, data = dfMC)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.5876 1.7217 -0.341 0.737
x4 0.2444 0.1614 1.514 0.148
x5 0.1257 0.1663 0.756 0.460
x6 0.1748 0.2598 0.673 0.510
```

How to solve? Either choose one, or combine them (e.g., take the average)

```
car::vif(lm(y~x4 + x5 + x6, dfMC))
x4  x5  x6
21.499973 21.834277 1.095182
```

Steps for Model Building

Step 1) Write down your hypotheses.

The selected independent variables should make sense in attempting to explain the dependent variable.

Use: logic / theory / your experience / your intuition.

Step 2) Check data, relationships, and assumptions

Plot all your variables. Also make scatterplots between pairs of variables.

Check correlations for linear relationship and possible multicollinearity

Check distribution of variables (Normally distributed? Bimodal?)

Check amount of missing data

Step 3) Use a systematic approach to building your model

Use an analysis plan. E.g., write down and test your hypotheses or plan and do stepwise regression, or a series of ANOVAs.

Step 4) Evaluate and Interpret your model

Correlation != Causation (especially in a "predictive" model)

Principle of parsimony: All things being equal, simpler models are usually better.