Statistical Modelling with Data

May 23 – June 02, 2023

Instructor: Qing (Leah) Li, Ph.D. Candidate at Cumming School of Medicine

qing.li2@uclagary.ca

Thank you Dr. Thuntida Ngamkham for contributing the contents

Thank you Dr. Qingrun Zhang and Dr. Quan Long for contributing some slides

Statistical Modelling with Data

- Topic 1: Statistical Modelling
  - Lecture 1: First-order models with quantitative independent variables
- Topic 2: Statistical Modelling with interactions (Assignment 1)
  - Lecture 2: Interaction effects, quantitative and qualitative variables
  - Lecture 3: Interaction effects and second-order models
- Topic 3: Statistical Model selection (Assignment 2)
  - Lecture 4: Model selection: Stepwise regression procedures
  - Lecture 5: Model selection: Forward and Backward selection procedures
- Topic 4: Statistical model diagnostics
  - Lecture 6: Multiple regression diagnostics: verify linearity, independence, and equal variance assumptions.
  - Lecture 7: Multiple regression diagnostics: verify normality assumptions and identify multicollinearity and outliers.
  - Lecture 8: Multiple regression diagnostics: data transformation
- Topic 5: Transfer learning
  - Lecture 9: Transfer-learning (Bonus): standing on the shoulders of giants.

Statistical Modelling with Data

- Topic 1: Statistical Modelling
  - Lecture 1: First-order models with quantitative independent variables
- Topic 2: Statistical Modelling with interactions (Assignment 1)
  - Lecture 2: Interaction effects, quantitative and qualitative variables
  - Lecture 3: Interaction effects and second-order models
- Topic 3: Statistical Model selection (Assignment 2)
  - Lecture 4: Model selection: Stepwise regression procedures
  - Lecture 5: Model selection: Forward and Backward selection procedures
- Topic 4: Statistical model diagnostics
  - Lecture 6: Multiple regression diagnostics: verify linearity, independence, and equal variance assumptions.
  - Lecture 7: Multiple regression diagnostics: verify normality assumptions and identify multicollinearity and outliers.
  - Lecture 8: Multiple regression diagnostics: data transformation
- Topic 5: Transfer learning
  - Lecture 9: Transfer-learning (Bonus): standing on the shoulders of giants.

Statistical Modelling with Data

Learning Outcomes: At the end of the course, participants will be able to

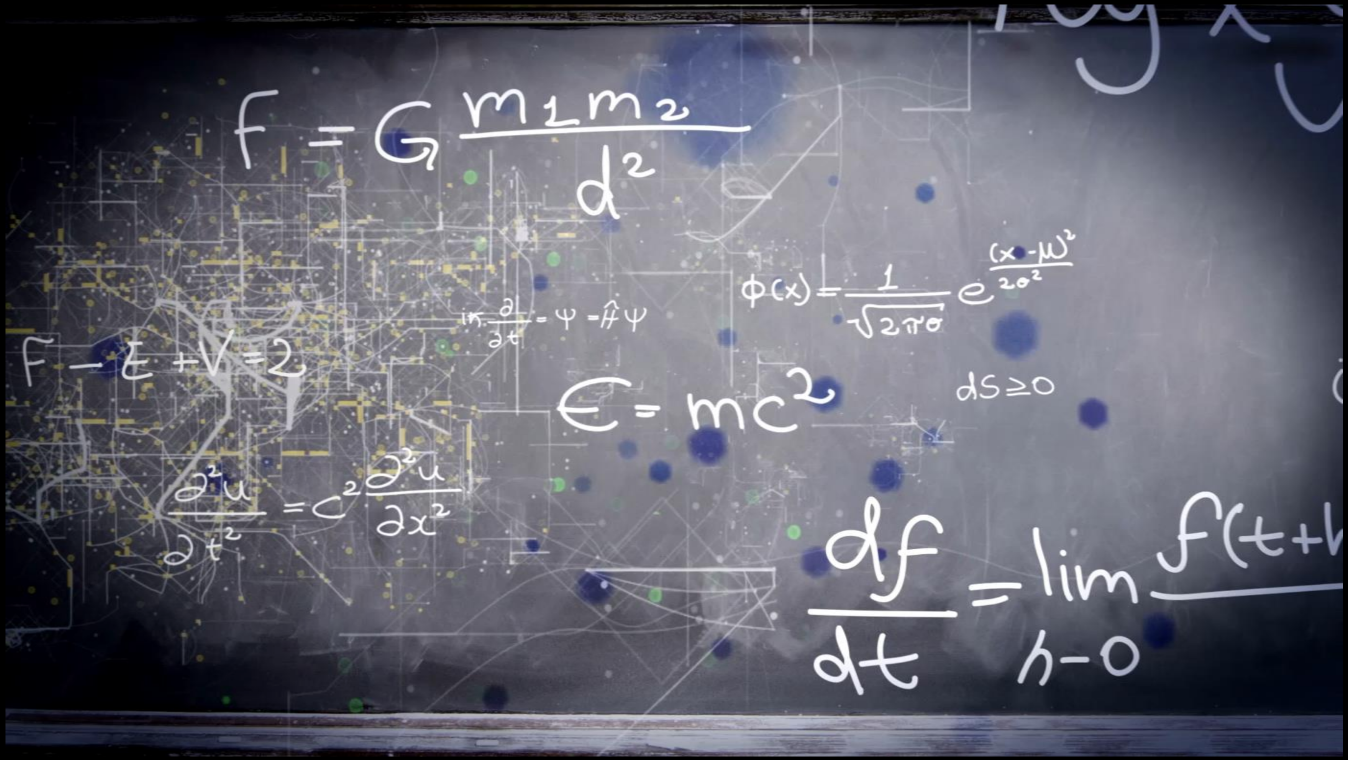
1. Model the multiple linear relationships between a response variable (Y) and all explanatory variables (both categorical and numerical variables) with interaction terms. Interpret model parameter estimates, construct confidence intervals for regression coefficients, evaluate model fits, and visualize correlations between a response variable (Y) and all explanatory variables (X) by graphs (scatter plot, residual plot) to assess model validity.
2. Predict the response variable at a certain level of the explanatory variables once the fit model exists.
3. Implement R-software and analyze statistical results for biomedical and other data.

- **Evaluations**
  1. Assignments will be posted on Slack (our communication tool with students).
  2. Students must attend 70% (6/9) of the sessions in order to receive the certificate and are encouraged to work on the assignments progressively throughout the course as the relevant material is covered.

Statistical Modelling with Data

- Supportive materials
  - Lectures slides (2023)
  - R code scripts (2023)
  - PDF (dated 2022)
  - Two Assignments (dated 2022)
- Slack channels
  - Recoding videos
  - Exercises
  - Course-documents

Lecture 2: Multiple Linear Regression Interaction effects, quantitative and qualitative variables



Quick recap of lecture 1

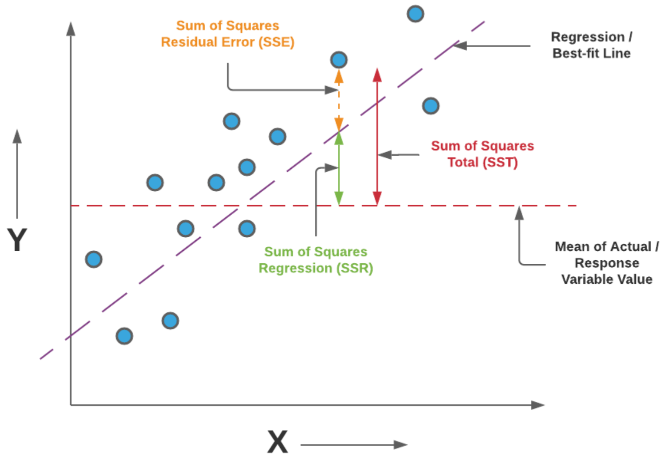
response, coefficients, predictors

- Statistics:
  - General linear model $g(y_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, $g(y_i) = y_i$
  - Generalized linear model $g(y_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, $g(y_i) = \ln(y_i), g(y_i) = logit(p) = \ln(\frac{p}{1-p})$
  - Least square method to estimate parameters

    The least squares estimates (**LSE**) $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$ are obtained by **minimizing the sum of the squared residuals:**

    $$\text{SSE}=\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_p X_{pi}))^2$$

  - Model utility (different null hypothesis H0)
    - F-test : full VS null — H0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
    - Partial F-test: full VS reduced — H0: $\beta_1 = 0$ and $\beta_2 = \beta_3 = \beta_4 \neq 0$
    - t-test: individual predictor — H0: $\beta_i = 0$
  - Model goodness of fit: R^2, R^2_{adj}, MSE
  - Model prediction
- Code:
  - lm(); glm(); summary(); coefficients(); confint(); anova(); predict();



The ANOVA table for Multiple Linear Regression

Source of Variation	DF	Sum of Squares	Mean Square	F-Statistic
Regression	p	SSR	MSR	MSR/MSE
Residual	n-p-1	SSE	MSE	
Total	n-1	SST		

An Interaction Model with Quantitative Predictors

- Standard linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- According to this model, if we increase *X*1 by one unit, then *Y* will increase by an average of β_1 units.
- Notice that the presence of *X*2 does not alter this statement-that is, regardless of the value of *X*2, a one-unit increase in *X*1 will lead to a β_1 unit increase in *Y* .
- The above equation is also known as additive model, investigating only the main effects of predictors. It assumes that the relationship between a given predictor variable and the response is independent of the other predictor variable.

An Interaction Model with Quantitative Predictors

- Family wellness = health + wealth + wife + husband + kids + interactions

Each predictor is known as term in statistics
Predictor = term = variable

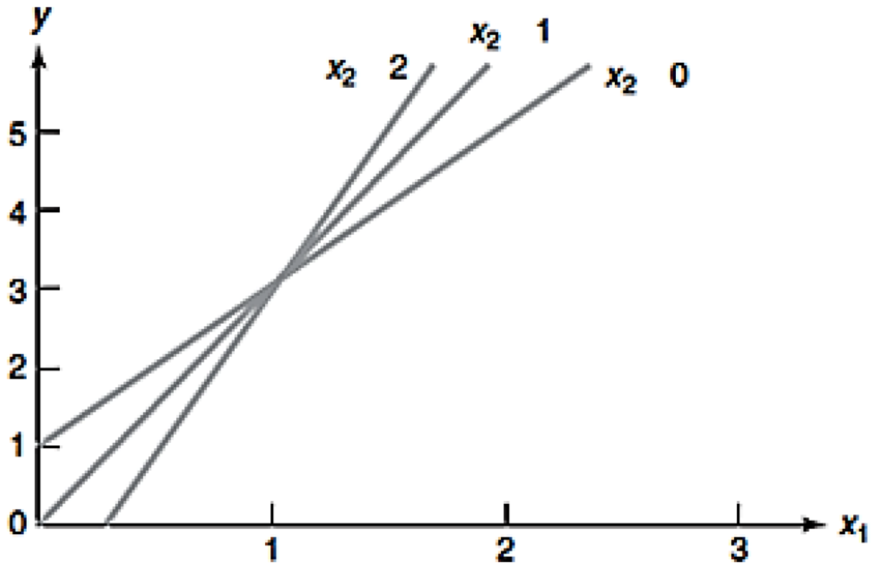
An Interaction Model with Quantitative Predictors

Interaction occurs when the influence of an independent variable on a dependent variable is not consistent across all independent variable values.
In this case, we need another model that will take into account this dependence.
Such a model includes the cross products of two or more *X*'s.
Hence, the interaction model for two variables looks like below:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

An example

Influence of the independent variable X1 on the dependent variable Y is not consistent across X2 values!



$$E(Y) = 1 + 2X_1 - X_2 + X_1X_2$$

For $X_2 = 0$:

$E(Y) = 1 + 2X_1 - (0) + X_1(0) = 1 + 2X_1 (slope = 2)$

For $X_2 = 1$:

$E(Y) = 1 + 2X_1 - (1) + X_1(1) = 3X_1 (slope = 3)$

For $X_2 = 2$:

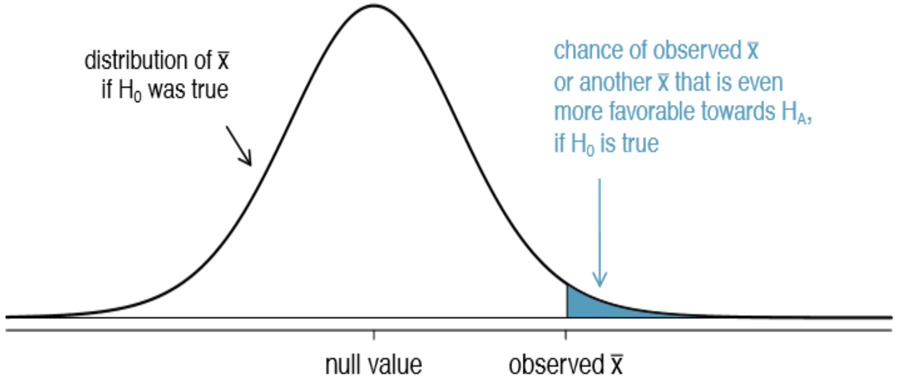
$E(Y) = 1 + 2X_1 - (2) + X_1(2) = -1 + 4X_1 (slope = 4)$

Note that the slope of each line is represented by slope$=\beta_1 + \beta_3 x_2 = 2 + x_2$. Thus, the effect on $E(Y)$ of a change in X_1 (i.e., the slope) now depends on the value of X_2. When this situation occurs, we say that X_1 and X_2 interact. Otherwise, the graph for 3 lines would be parallel. The cross-product term, X_1X_2, is called an interaction term, and the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$ is called **an interaction model with two quantitative variables.**

Testing for Interaction in Multiple Regression

For testing an interaction term in regression model, we use the Individual Coefficients Test (t-test) method.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$



$$H_0 \quad : \beta_i = 0$$
$$H_a \quad : \beta_i \neq 0 \ (\mathrm{i} = 1,2,\ldots,\mathrm{p})$$

$$t_{cal} = \frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)} \text{ which has df} = \mathrm{n} - \mathrm{p} \text{ degree of freedom}$$

If the p-value for the t test statistic is extremely little (typically less than 0.05), it implies that our test statistic is unlikely to come from a null distribution. As a result, we can reject the null hypothesis with a little chance (alpha) of making mistakes. This also means that we accept the alternative hypothesis, and that the interaction term is critical in our statistical model.

Testing for Interaction in Multiple Regression

+a	include this variable a
a:b	interaction between two variables, a and b.
a*b	equivalent to a+b+a:b
(a+b)^2	equivalent to a+b+a:b
(a+b+c)^2	a+b+c+a:b+a:c+b:c

```
> interacmodel<-lm(sale~tv+radio+tv:radio, data=Advertising)
> summary(interacmodel)

Call:
lm(formula = sale ~ tv + radio + tv:radio, data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.3366 -0.4028  0.1831  0.5948  1.5246 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
tv          1.910e-02  1.504e-03  12.699   <2e-16 ***
radio       2.886e-02  8.905e-03   3.241   0.0014 ** 
tv:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,	Adjusted R-squared:  0.9673 
F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
> interacmodel1<-lm(sale~tv*radio, data=Advertising)
> summary(interacmodel1)

Call:
lm(formula = sale ~ tv * radio, data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.3366 -0.4028  0.1831  0.5948  1.5246 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
tv          1.910e-02  1.504e-03  12.699   <2e-16 ***
radio       2.886e-02  8.905e-03   3.241   0.0014 ** 
tv:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,	Adjusted R-squared:  0.9673 
F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
> interacmodel2<-lm(sale~(tv + radio)^2, data=Advertising)
> summary(interacmodel2)

Call:
lm(formula = sale ~ (tv + radio)^2, data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.3366 -0.4028  0.1831  0.5948  1.5246 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
tv          1.910e-02  1.504e-03  12.699   <2e-16 ***
radio       2.886e-02  8.905e-03   3.241   0.0014 ** 
tv:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,	Adjusted R-squared:  0.9673 
F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

Interpreting Coefficients of Predictor Variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

$$\downarrow$$

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \phi X_1 + \beta_2 X_2 + \epsilon \\ &\mathit{where} \\ \phi &= \beta_1 + \beta_3 X_2. \end{aligned}$$

- Since ϕ changes with X_2, the effect of X_1 on Y is no longer constant: adjusting X_2 will change the impact of X_1 on Y

Interpreting Coefficients of Predictor Variables

- Advertising data

```
> reduced<-lm(sale~tv+radio, data=Advertising)
> newdata = data.frame(tv=200, radio=20)
> predict(reduced,newdata,interval="predict")
       fit      lwr     upr
1 15.83195 12.5042 19.1597
```

$$
\begin{aligned}
Sale &= \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 (TV * radio) + \epsilon \\
&= \beta_0 + (\beta_1 + \beta_3 radio) * TV + \beta_2 radio + \epsilon
\end{aligned}
$$

We can interpret the coefficient $\beta_{1+}\beta_3 radio$ as: spending additional 1,000 dollars on TV advertising leads to an *increase* in sales by approximately $\beta_{1+}\beta_3 radio$ units.

Interpreting Coefficients of Predictor Variables

```
> interacmodel2<-lm(sale~(tv + radio)^2, data=Advertising)
> summary(interacmodel2)

Call:
lm(formula = sale ~ (tv + radio)^2, data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.3366 -0.4028  0.1831  0.5948  1.5246 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
tv          1.910e-02  1.504e-03  12.699   <2e-16 ***
radio       2.886e-02  8.905e-03   3.241   0.0014 ** 
tv:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,	Adjusted R-squared:  0.9673 
F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

Is it necessary to include the interaction term or not?

The results from the output strongly suggest that the model that includes the interaction term is superior to the model that contains only main effects.

Sale = 6.75 + (1.91E-2 + 1.086E-3 radio) x TV + 2.886E-2 radio + ε

The coefficient estimates in the output suggest that an increase in TV advertising of 1,000 dollars is associated with increased sales of $(\beta_{1+}\beta_3 radio) \times 1000 = 19 + 1.1 radio$ units. And an increase in radio advertising of 1,000 dollars will be associated with an increase in sales of $(\beta_{2+}\beta_3 TV) \times 1,000 = 29 + 1.1 TV$) units.

One Health at UCalgary | UNIVERSITY OF CALGARY

Interpreting Coefficients of Predictor Variables

- In this example, the p-values associated with TV, radio, and the interaction term all are statistically significant and so it is obvious that all three variables should be included in the model. However, it is sometimes the case that an interaction term has a very small p-value, but the associated main effects (in this case, TV and radio) do not.
- The hierarchical principle states that if we include an interaction in a model, we should also include the main effects, even if the p-values associated with principle their coefficients are not significant.

Caution

- If the interaction between $X1$ and $X2$ seems important, then we should include both $X1$ and $X2$ in the model even if their coefficient estimates have large p-values.
- The rationale for this principle is that if $X1 \times X2$ is related to the response, then whether or not the coefficients of X1 or X2 are exactly zero is of little interest. Also, $X1 \times X2$ is typically correlated with $X1$ and $X2$, and so leaving them out tends to alter the meaning of the interaction.

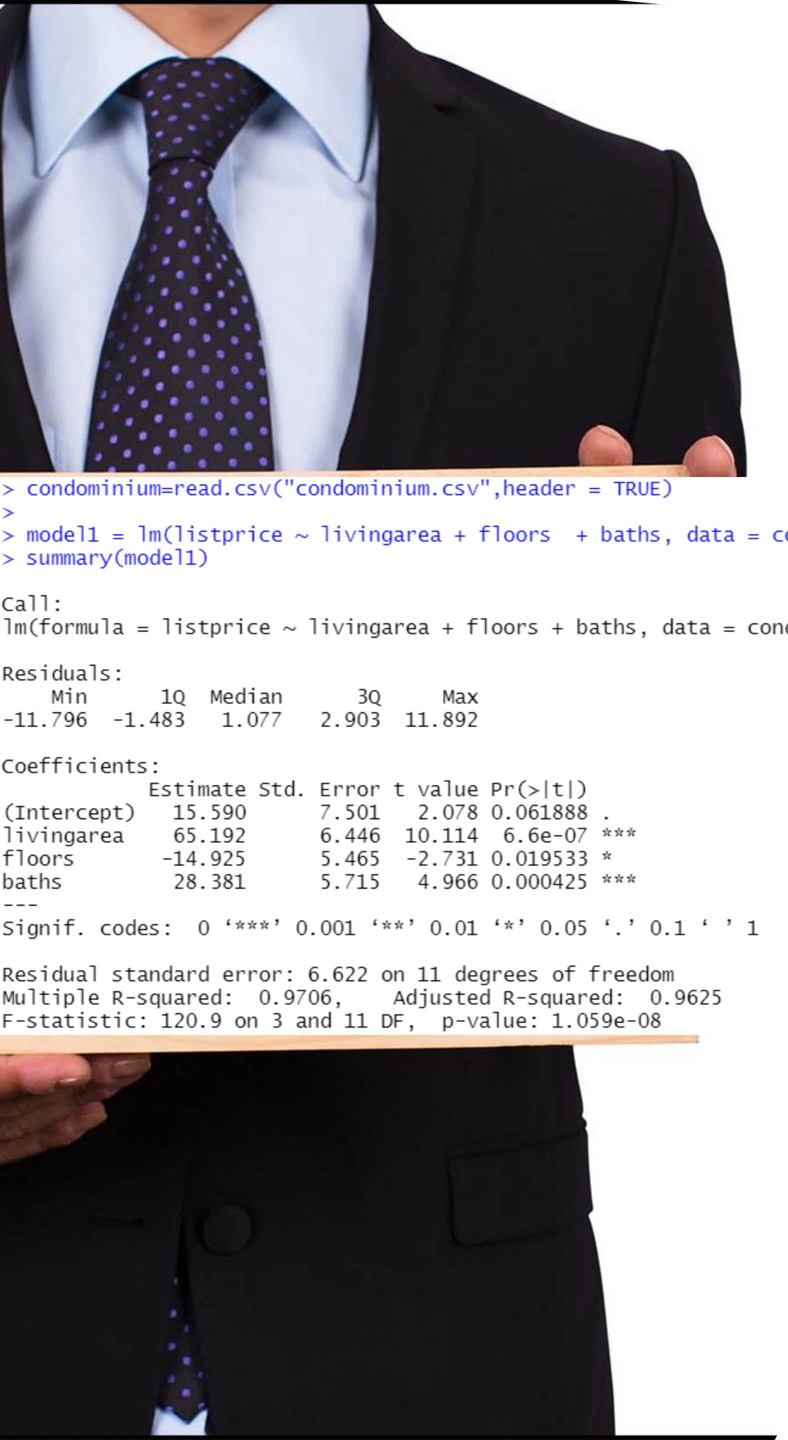
In class Practice Problem 4

From the condominium problem, do the data provide sufficient evidence to indicate that the interaction term need to be added in the model? If you had to compare additive models with the interaction model, which model would you choose? Explain.

```
> condominium=read.csv("condominium.csv",header = TRUE)
> full<-lm(listprice ~ livingarea + floors + bedrooms + baths, data=condominium)
> fit <-lm(listprice ~ livingarea + floors+ baths , data=condominium)
>
> summary(full)$adj.r.squared
[1] 0.9602536
> sigma(full)
[1] 6.819782
>
> summary(fit)$adj.r.squared
[1] 0.9625232
> sigma(fit)
[1] 6.622212
>
> newdata = data.frame(livingarea=2, floors=3, baths=3)
> predict(fit,newdata,interval="predict") #95% prediction interval
       fit     lwr      upr
1 186.3426 166.194 206.4911
```

L2.R

In class Practice Problem 4



```
> condominium=read.csv("condominium.csv",header = TRUE)
>
> model1 = lm(listprice ~ livingarea + floors  + baths, data = condominium)
> summary(model1)

Call:
lm(formula = listprice ~ livingarea + floors + baths, data = condominium)

Residuals:
    Min      1Q  Median      3Q     Max
-11.796  -1.483   1.077   2.903  11.892

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.590      7.501   2.078 0.061888 .
livingarea    65.192      6.446  10.114  6.6e-07 ***
floors       -14.925      5.465  -2.731 0.019533 *
baths         28.381      5.715   4.966 0.000425 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.622 on 11 degrees of freedom
Multiple R-squared:  0.9706,	Adjusted R-squared:  0.9625
F-statistic: 120.9 on 3 and 11 DF,  p-value: 1.059e-08
```

```
> model2 = lm(listprice ~ livingarea + floors  + baths + livingarea:floors+ livingarea:baths+ floors:baths, data = condominium)
> summary(model2)

Call:
lm(formula = listprice ~ livingarea + floors + baths + livingarea:floors +
    livingarea:baths + floors:baths, data = condominium)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2701  -2.0116  -0.4466   3.8389   6.2406

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         40.502     77.021   0.526    0.613
livingarea          27.339     44.761   0.611    0.558
floors             -19.319    109.400  -0.177    0.864
baths               28.625     24.946   1.147    0.284
livingarea:floors   13.143     25.315   0.519    0.618
livingarea:baths    10.209     31.157   0.328    0.752
floors:baths        -8.062     37.341  -0.216    0.834

Residual standard error: 6.611 on 8 degrees of freedom
Multiple R-squared:  0.9787,	Adjusted R-squared:  0.9626
F-statistic: 61.13 on 6 and 8 DF,  p-value: 3.008e-06
```

Data does not provide sufficient evidence to indicate that the interaction term need to be added in the model.

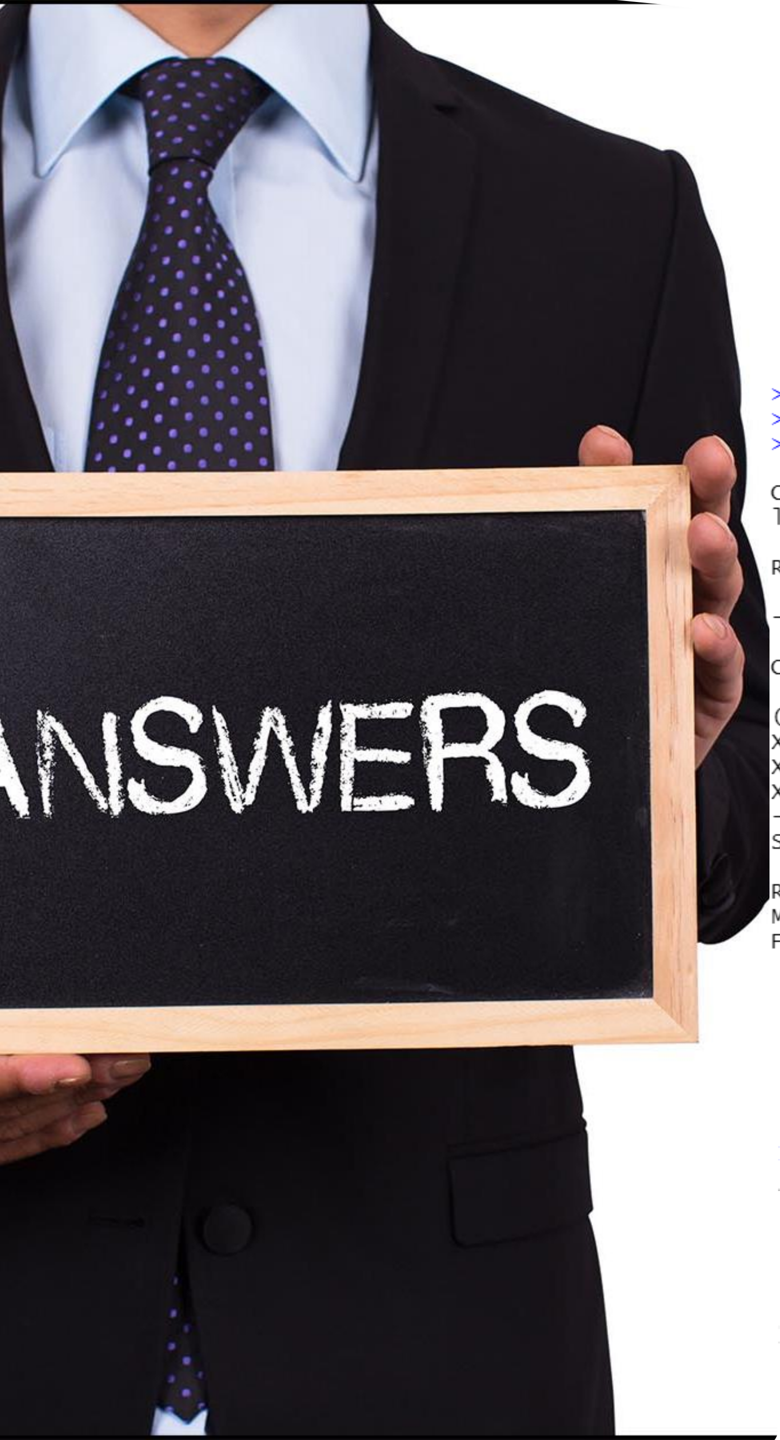
In class Practice Problem 5

Data on last year's sale (Y in 100,000s dollars) for 40 sales districts (sales.csv). This file also contains

- promotional expenditures ($X1$: in 1,000s dollars),
- the number of active accounts ($X2$),
- the number of competing brands ($X3$) and
- the district potential ($X4$, coded) for each of the districts (OMIT THIS VARIABLE FOR NOW)

1. Find the best fit additive model to predict sales using some or all of the variables $X1, X2, X3$ only.
2. Find the best fit model with interaction terms (if needed) using some or all of the variables $X1, X2, X3$
3. Which model would you choose? Explain.
4. Once you obtain the best fit model, interpret the regression coefficient for $X3$ (Hint: it will interact with another variable).

L2.R



One Health at UCalgary

UNIVERSITY OF CALGARY

In class Practice Problem 5

```
> sales=read.csv("sales.csv",header = TRUE)
> model1 = lm(formula = Y ~ X1 + X2 + X3, data=sales)
> summary(model1)

Call:
lm(formula = Y ~ X1 + X2 + X3, data = sales)

Residuals:
     Min       1Q   Median       3Q      Max
-106.803   -6.726   -1.967    7.072   81.964

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 162.2269    31.0376   5.227 7.50e-06 ***
X1            2.0192     2.5763   0.784    0.438
X2            3.4568     0.3426  10.088 4.91e-12 ***
X3          -19.4589     1.8054 -10.778 8.08e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.35 on 36 degrees of freedom
Multiple R-squared:  0.9175,    Adjusted R-squared:  0.9106
F-statistic: 133.4 on 3 and 36 DF,  p-value: < 2.2e-16
```

```
> model2 = lm(formula = Y ~ X2 + X3, data=sales)
> summary(model2)

Call:
lm(formula = Y ~ X2 + X3, data = sales)

Residuals:
     Min       1Q   Median       3Q      Max
-109.096   -5.888   -3.440    8.780   83.982

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 172.4595    28.0109   6.157 3.85e-07 ***
X2            3.5011     0.3362  10.414 1.50e-12 ***
X3          -19.7308     1.7625 -11.195 1.94e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.22 on 37 degrees of freedom
Multiple R-squared:  0.9161,    Adjusted R-squared:  0.9115
F-statistic: 201.9 on 2 and 37 DF,  p-value: < 2.2e-16
```

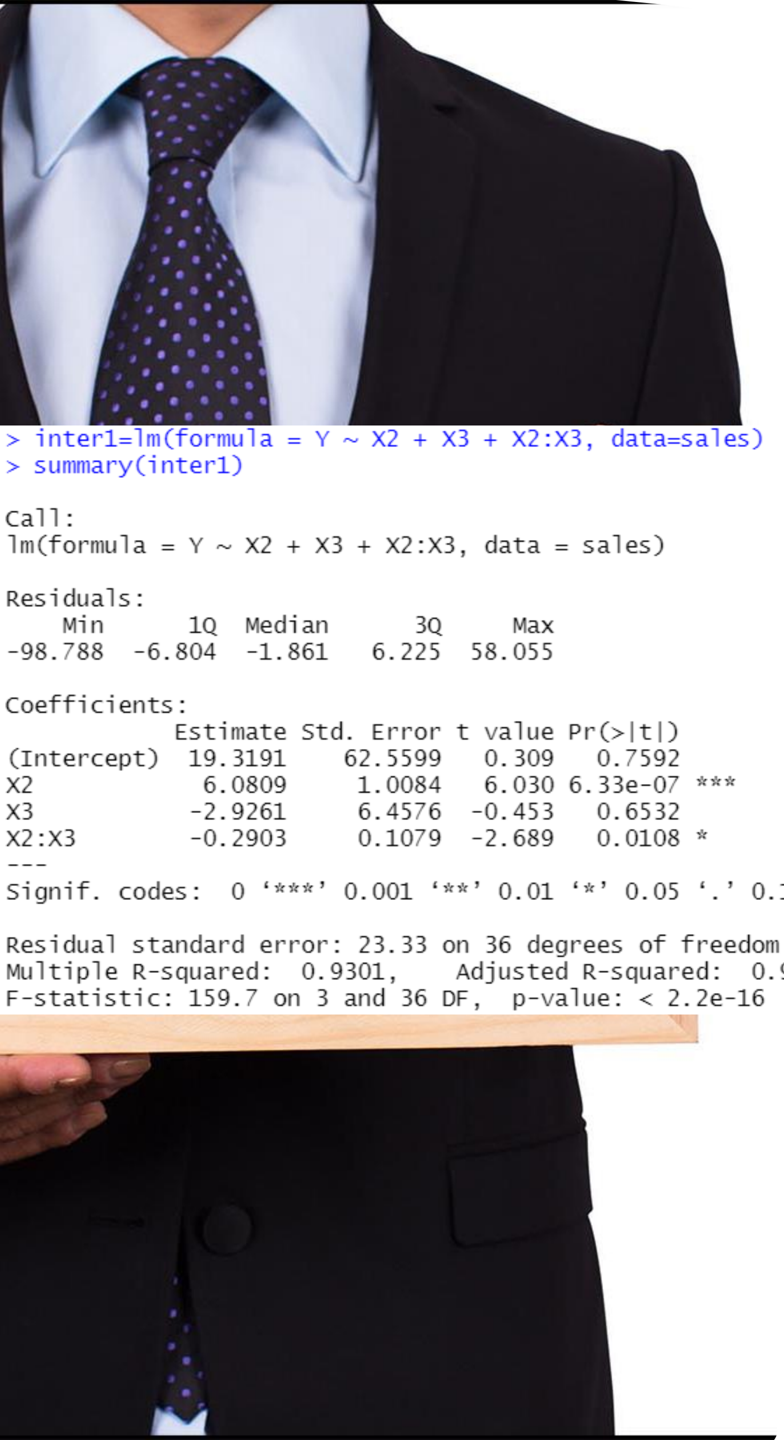
Partial F-test

H0: coefficient of X1 =0

```
> anova(model2,model1)
Analysis of Variance Table

Model 1: Y ~ X2 + X3
Model 2: Y ~ X1 + X2 + X3
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     37 23532
2     36 23137  1    394.79 0.6143 0.4383
```

P value is not significant, H0 cannot be rejected. So, the coefficient for X1 can be considered as 0 and we exclude X1 from the model.

In class Practice Problem 5



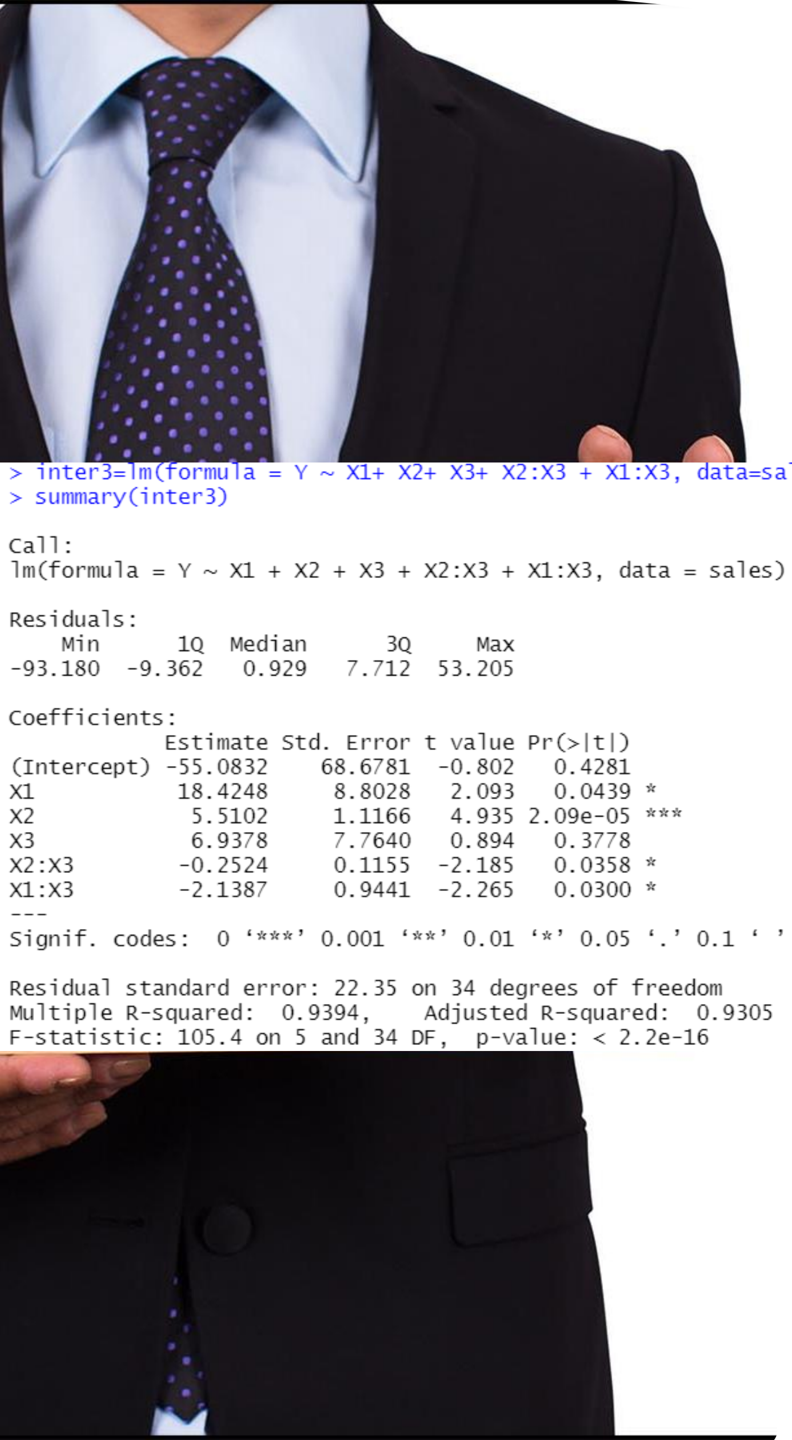
```
> inter1=lm(formula = Y ~ X2 + X3 + X2:X3, data=sales)
> summary(inter1)

Call:
lm(formula = Y ~ X2 + X3 + X2:X3, data = sales)

Residuals:
    Min      1Q  Median      3Q     Max 
-98.788  -6.804  -1.861   6.225  58.055 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  19.3191    62.5599   0.309   0.7592    
X2            6.0809     1.0084   6.030 6.33e-07 ***
X3           -2.9261     6.4576  -0.453   0.6532    
X2:X3        -0.2903     0.1079  -2.689   0.0108 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.33 on 36 degrees of freedom
Multiple R-squared:  0.9301,	Adjusted R-squared:  0.9243 
F-statistic: 159.7 on 3 and 36 DF,  p-value: < 2.2e-16
```

```
> inter2=lm(formula = Y ~ (X1+X2+X3)^2, data=sales)
> summary(inter2)

Call:
lm(formula = Y ~ (X1 + X2 + X3)^2, data = sales)

Residuals:
    Min      1Q  Median      3Q     Max 
-93.253  -9.208   0.852   6.606  51.455 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -22.0421    91.7901  -0.240  0.81171   
X1           11.9200    14.7995   0.805  0.42633   
X2            4.8325     1.6708   2.892  0.00672 **
X3            7.0945     7.8501   0.904  0.37268   
X1:X2         0.1193     0.2169   0.550  0.58607   
X1:X3        -2.1302     0.9540  -2.233  0.03246 * 
X2:X3        -0.2495     0.1168  -2.136  0.04021 * 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.59 on 33 degrees of freedom
Multiple R-squared:   0.94,	Adjusted R-squared:  0.929 
F-statistic: 86.09 on 6 and 33 DF,  p-value: < 2.2e-16
```

```
> inter3=lm(formula = Y ~ X1+ X2+ X3+ X2:X3 + X1:X3, data=sales)
> summary(inter3)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X2:X3 + X1:X3, data = sales)

Residuals:
    Min      1Q  Median      3Q     Max 
-93.180  -9.362   0.929   7.712  53.205 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -55.0832    68.6781  -0.802   0.4281    
X1           18.4248     8.8028   2.093   0.0439 *  
X2            5.5102     1.1166   4.935 2.09e-05 ***
X3            6.9378     7.7640   0.894   0.3778    
X2:X3        -0.2524     0.1155  -2.185   0.0358 *  
X1:X3        -2.1387     0.9441  -2.265   0.0300 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.35 on 34 degrees of freedom
Multiple R-squared:  0.9394,	Adjusted R-squared:  0.9305 
F-statistic: 105.4 on 5 and 34 DF,  p-value: < 2.2e-16
```

ANOVA test suggested X1 can be exclude due to its insignificant p-values
However, interactions between X1 and X3 are significant and indicate it’s not appropriate to exclude X1.
We suggest you include all predictors when you build your first interaction model for your data.

In class Practice Problem 5



```
> inter3=lm(formula = Y ~ X1+ X2+ X3+ X2:X3 + X1:X3, data=sales)
> summary(inter3)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X2:X3 + X1:X3, data = sales)

Residuals:
    Min      1Q  Median      3Q     Max 
-93.180  -9.362   0.929   7.712  53.205 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -55.0832    68.6781  -0.802   0.4281    
X1           18.4248     8.8028   2.093   0.0439 *  
X2            5.5102     1.1166   4.935 2.09e-05 ***
X3            6.9378     7.7640   0.894   0.3778    
X2:X3        -0.2524     0.1155  -2.185   0.0358 *  
X1:X3        -2.1387     0.9441  -2.265   0.0300 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.35 on 34 degrees of freedom
Multiple R-squared:  0.9394,	Adjusted R-squared:  0.9305 
F-statistic: 105.4 on 5 and 34 DF,  p-value: < 2.2e-16
```

The model I chose is inter3: Y ~ X1+ X2+ X3+ X2:X3 + X1:X3
The coeficiente of X3 is : (-2.1387 x X1-0.2524 x X2+18.4248).

The coefficient estimates in the output suggest that an increase in the number of competing brands of 1 unit is associated with increased sales of (-2.1387 x promotional expenditures – 0.2524 x active accounts + 18.4248) x 1 units.

However, promotional expenditures and activate accounts are both positive and in scale of thousands. Therefore, increase of brands of 1 unit is most likely to reduce sales.

In this example, the main effect of X3 is not significantly associated with sales. However, its interactions with X1 and X2 are significantly associated with sales. Therefore, we'd better include X1 in our model.

Coffee break

Multiple Regression with Qualitative (Dummy) Variable Models

	Categorical	**Quantitative**
Definition	*Take on names or labels*	*Take on numeric values*
Examples	Marital Status	Height
	Smoking Status	Population Size
	Eye Color	Square Footage
	Level of Education	Class Size

Nominal

The categories of a nominal variable have no inherent or natural order

Ordinal

The categories of a nominal variable have no inherent or natural order

- Multiple regression models can also be written to include qualitative (or categorical) independent variables.
- Qualitative variables, unlike quantitative variables, cannot be measured on a numerical scale. Therefore, we must code the values of the qualitative variable (called levels) as numbers before we can fit them in the model.
- It's easy to code ordinal variable as they have an inherent order. But how about nominal variables? Does it make sense to code 0 for blue eye colour, 1 for green eye colour and 2 for hazel eye colour?

Multiple Regression with Qualitative (Dummy) Variable Models

- Because categorical predictor variables cannot be entered directly into a regression model and be meaningfully interpreted, some other methods of dealing with information of this type must be developed.
- In Generally, a categorical variable with k levels will be transformed into k −1 variables with two levels. For example, if a categorical variable had six levels, then five dichotomous variables could be constructed that would contain the same information as the single categorical variable.
- The process of creating dichotomous variables from categorical variables is called dummy coding. These dichotomous variables are called dummy variables. The simplest case of dummy coding is when a categorical variable has two levels by assigning zero and one to the variable.

Smoking status	X1
Smoker	1
non-smoker	0

Eyes colour	X1	X2
Blue	1	0
Green	0	1
Hazel	0	0

Multiple Regression with Qualitative (Dummy) Variable Models

```
> head(credit)
  number  Income Limit Rating Cards Age Education Gender Student Married Ethnicity Balance
1      1  14.891  3606    283     2  34        11   Male      No     Yes Caucasian     333
2      2 106.025  6645    483     3  82        15 Female     Yes     Yes     Asian     903
3      3 104.593  7075    514     4  71        11   Male      No      No     Asian     580
4      4 148.924  9504    681     3  36        11 Female      No      No     Asian     964
5      5  55.882  4897    357     2  68        16   Male      No     Yes Caucasian     331
6      6  80.180  8047    569     4  77        10   Male      No      No Caucasian    1151
```

- The Credit data set records balance (average credit card debt for a number of individuals) as well as several quantitative predictors: age, cards (number of credit cards), education (years of education), income (in thousands of dollars), limit (credit limit), and rating (credit rating).
- In addition to these quantitative variables, we also have four qualitative variables: gender, student (student status), status (marital status), and ethnicity (Caucasian, African American or Asian). Data are provided in credit.csv file.
- Suppose that we wish to investigate differences in credit card balance between males and females. Based on the gender variable, we can create one dummy variable (2-1=1) with 0 as male and 1 as female.

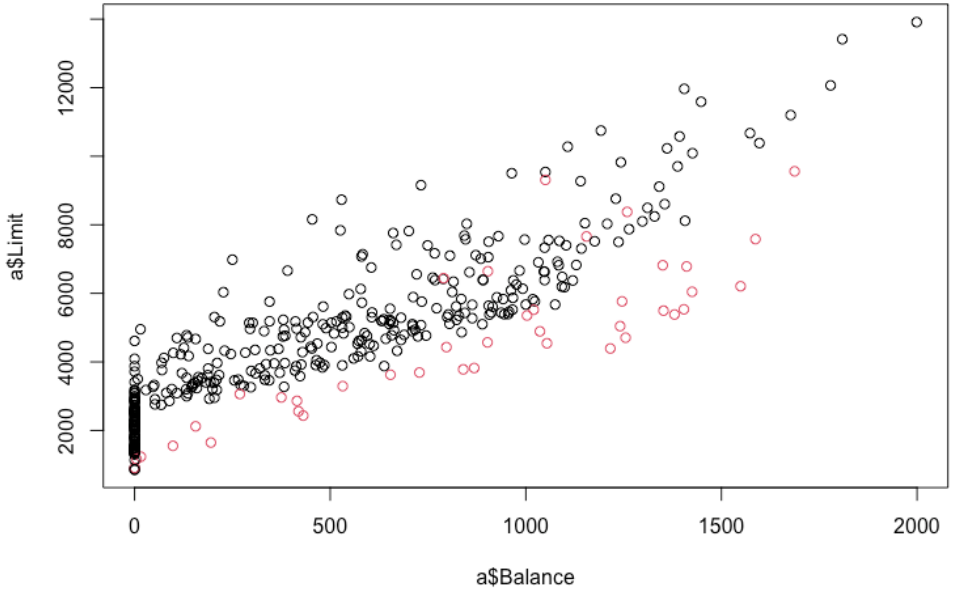
Multiple Regression with Qualitative (Dummy) Variable Models

Gender	X1
Male	1
Female	0

Type equation here.

$$Y_i \quad = \beta_0 + \beta_1 X_i + \epsilon$$

$$balance_i \quad = \begin{cases} \beta_0 + \beta_1 + \epsilon & \text{if } i^{th} \text{ person is female} \\ \beta_0 + \epsilon & \text{if } i^{th} \text{ person is male} \end{cases}$$



Multiple Regression with Qualitative (Dummy) Variable Models

```
> dummymodel1<-lm(Balance~Gender,data=credit)
> summary(dummymodel1)

Call:
lm(formula = Balance ~ Gender, data = credit)

Residuals:
    Min      1Q  Median      3Q     Max 
-529.54 -455.35  -60.17  334.71 1489.20 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    509.80      33.13  15.389   <2e-16 ***
GenderFemale    19.73      46.05   0.429    0.669    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 460.2 on 398 degrees of freedom
Multiple R-squared:  0.0004611,	Adjusted R-squared:  -0.00205 
F-statistic: 0.1836 on 1 and 398 DF,  p-value: 0.6685
```

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon$$

$$balance_i = \begin{cases} \beta_0 + \beta_1 + \epsilon & \text{if } i^{th} \text{ person is female} \\ \beta_0 + \epsilon & \text{if } i^{th} \text{ person is male} \end{cases}$$

```
> dummymodel<-lm(Balance~factor(Gender),data=credit)
> summary(dummymodel)

Call:
lm(formula = Balance ~ factor(Gender), data = credit)

Residuals:
    Min      1Q  Median      3Q     Max 
-529.54 -455.35  -60.17  334.71 1489.20 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)            509.80      33.13  15.389   <2e-16 ***
factor(Gender)Female    19.73      46.05   0.429    0.669    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 460.2 on 398 degrees of freedom
Multiple R-squared:  0.0004611,	Adjusted R-squared:  -0.00205 
F-statistic: 0.1836 on 1 and 398 DF,  p-value: 0.6685
```

factor() : command will make sure that R knows that your variable is categorical. This is especially useful if your categories are indicated by integers, otherwise function lm() might interpret the variable as continuous.

Interpreting coefficients of predictor variable

β_0 can be interpreted as the average credit card balance among males,

β_1 as the average difference in credit card balance between females and males.

$\beta_0 + \beta_1$ can be interpreted as the average credit card balance among females.

$$\hat{Y}_i = 509.80 + 19.73X_i$$

$$balance_i = \begin{cases} 509.80 + 19.73 = 529.53 & \text{if } i^{th} \text{ person is female} \\ 509.80 & \text{if } i^{th} \text{ person is male} \end{cases}$$

- From the output, the coefficient estimates and other information associated with the model are provided. The average credit card debt for males is estimated to be 509.80 dollars whereas females are estimated to carry 19.73 in additional debt for a total of 509.80 + 19.73 = 529.53 dollars.
- However, we notice that the p-value (0.669) for the dummy variable is very high. This indicates that there is no statistical evidence of a difference in average credit card balance between the genders.

One Health at UCalgary

UNIVERSITY OF CALGARY

In class Practice Problem 6

Suppose that we wish to investigate differences in credit card balance between marital status. Based on the Married variable, we can create a dummy variable which 0 is NO and 1 is Yes.

- Create a simple linear regression model to predict the credit card balance by using the Married variable.
- How much is the average credit card debt for an unmarried person.
- What is the difference in debt between a married and single person.

Ignore the individual t-test output.

L2.R

In class Practice Problem 6

```
> dummymodel2<-lm(Balance~factor(Married),data=credit)
> summary(dummymodel2)

Call:
lm(formula = Balance ~ factor(Married), data = credit)

Residuals:
    Min      1Q  Median      3Q     Max 
-523.29 -451.03  -60.12  345.06 1481.06 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         523.290     36.974  14.153   <2e-16 ***
factor(Married)Yes   -5.347     47.244  -0.113     0.91    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 460.3 on 398 degrees of freedom
Multiple R-squared:  3.219e-05,	Adjusted R-squared:  -0.00248 
F-statistic: 0.01281 on 1 and 398 DF,  p-value: 0.9099
```

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon \\ balance_i &= \begin{cases} \beta_0 + \beta_1 + \epsilon \\ \beta_0 + \epsilon \end{cases} \end{aligned}$$

The average credit card balance for an unmarried person is 523.29. The average is 523.290-5.347=517.943 for a married person.

L2.R

Dummy Coding with three levels

For example, there is always a certain curiosity and controversy surrounding professor' salaries and whether they are overpaid or not paid enough. A university would like to study the effects of ranks and department on the salaries. 30 observations were randomly chosen from 3 different departments. The data are provided in salary.csv data file.

Dummy Coding with three levels:

dummy variable:

Deaprtment	Biology	Business
Family Studies	0	0
Biology	1	0
Business	0	1

Dummy Coding with three levels

Dept= Department (1 =Family Studies, 2 =Biology, 3 = Business).

```
> salary=read.csv("salary.csv",header = TRUE)
> head(salary)
  salary gender rank dept year merit
1     38      0    3    1    0  1.47
2     58      1    2    2    8  4.38
3     80      1    3    2    9  3.65
4     30      1    1    1    0  1.64
5     50      1    1    3    0  2.54
6     49      1    1    3    1  2.06
```

! Coding variable as qualitative (categorical) or quantitative (continuous) leads to different regression results!

```
> dummymodel<-lm(salary~factor(rank),data=salary)
> summary(dummymodel)

Call:
lm(formula = salary ~ factor(rank), data = salary)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.875  -5.799   0.000   5.353  23.125 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     42.000      2.259  18.593  < 2e-16 ***
factor(rank)2   10.571      4.005   2.640 0.013613 *  
factor(rank)3   14.875      3.830   3.884 0.000602 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.749 on 27 degrees of freedom
Multiple R-squared:  0.3881,	Adjusted R-squared:  0.3428 
F-statistic: 8.563 on 2 and 27 DF,  p-value: 0.001319
```

```
> dummymodel1<-lm(salary~rank,data=salary)
> summary(dummymodel1)

Call:
lm(formula = salary ~ rank, data = salary)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.9017  -5.2168   0.1139   5.8639  22.0983 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   34.871      3.684   9.466 3.19e-10 ***
rank           7.677      1.882   4.080 0.000339 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.697 on 28 degrees of freedom
Multiple R-squared:  0.3729,	Adjusted R-squared:  0.3505 
F-statistic: 16.65 on 1 and 28 DF,  p-value: 0.0003389
```

Interpreting coefficients of predictor variables

```
> dummymodel<-lm(salary~factor(rank),data=salary)
> summary(dummymodel)

Call:
lm(formula = salary ~ factor(rank), data = salary)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.875  -5.799   0.000   5.353  23.125 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     42.000      2.259  18.593  < 2e-16 ***
factor(rank)2   10.571      4.005   2.640 0.013613 *  
factor(rank)3   14.875      3.830   3.884 0.000602 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.749 on 27 degrees of freedom
Multiple R-squared:  0.3881,	Adjusted R-squared:  0.3428 
F-statistic: 8.563 on 2 and 27 DF,  p-value: 0.001319
```

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon$$

$$salary_i = \begin{cases} \beta_0 + \beta_1 + \epsilon & \text{if } i^{th} \text{ person is ranked as Associate Prof} \\ \beta_0 + \beta_2 + \epsilon & \text{if } i^{th} \text{ person is ranked as Full Prof} \\ \beta_0 + \epsilon & \text{if } i^{th} \text{ person is ranked as Assistant Prof} \end{cases}$$

β_0 can be interpreted as the average salary for Assistant Professor position ,

β_1 as the difference in average salary between Associate Professor and Assistant Professor.

β_2 as the difference in average salary between Full Professor and Assistant Professor.

$\beta_0 + \beta_1$ can be interpreted as the average salary for Associate Professor position .

$\beta_0 + \beta_2$ can be interpreted as the average salary for Full Professor position .

In class Practice Problem 7

There is always a certain curiosity and controversy surrounding professors’ salaries and whether they are overpaid or not paid enough. A university would like to study the effects of ranks and departments on salaries. 30 observations were randomly chosen from 3 different departments. The data are provided in the salary.csv data file. Dept= Department (1=Family studies, 2=Biology, 3=Business).

1. Instead of the rank variable, practice how to interpret the department variable.

2. Can you also try to interpret rank and department together?

L2.R

In class Practice Problem 7

```
> dummymodel1<-lm(salary~factor(dept),data=salary)
> summary(dummymodel1)

Call:
lm(formula = salary ~ factor(dept), data = salary)

Residuals:
    Min      1Q  Median      3Q     Max 
-12.250  -6.838  -3.925   4.662  30.000 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      42.250      2.788  15.154 1.01e-14 ***
factor(dept)2     7.750      4.408   1.758  0.09008 .  
factor(dept)3    12.350      4.135   2.986  0.00594 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.658 on 27 degrees of freedom
Multiple R-squared:  0.2543,	Adjusted R-squared:  0.199 
F-statistic: 4.603 on 2 and 27 DF,  p-value: 0.01905
```

Dummy Coding with three levels:

dummy variable:

Deaprtment	Biology	Business
Family Studies	0	0
Biology	1	0
Business	0	1

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

Average salary for professor from Family studies (1) is 42.25
Average salary for professor from biology (2) is 42.25+7.75=50
Average salary for professor from biology (2) is 42.25+12.35=54.6

L2.R

In class Practice Problem 7

```
> dummymodel2<-lm(salary~factor(rank) + factor(dept),data=salary)
> summary(dummymodel2)

Call:
lm(formula = salary ~ factor(rank) + factor(dept), data = salary)

Residuals:
    Min      1Q  Median      3Q     Max
-11.243  -3.333  -0.049   2.350  20.256

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      34.049      2.308  14.754 7.62e-14 ***
factor(rank)2    13.208      2.983   4.427 0.000164 ***
factor(rank)3    15.194      2.797   5.433 1.22e-05 ***
factor(dept)2    10.502      2.972   3.533 0.001624 **
factor(dept)3    13.351      2.752   4.851 5.48e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.368 on 25 degrees of freedom
Multiple R-squared:  0.6998,    Adjusted R-squared:  0.6518
F-statistic: 14.57 on 4 and 25 DF,  p-value: 2.856e-06
```

	x1	x2
Assistant	0	0
Associate	1	0
Full	0	1

	x3	x4
Family studies	0	0
Biology	1	0
Business	0	1

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon$$

$$salary_i = \begin{cases} \beta_0 + \beta_1 + \epsilon & \text{if } i^{th} \text{person is ranked as Associate Prof and is from Family Studies dept} \\ \beta_0 + \beta_2 + \epsilon & \text{if } i^{th} \text{person is ranked as Full Prof and is from Family Studies dept} \\ \beta_0 + \beta_3 + \epsilon & \text{if } i^{th} \text{person is ranked as Assistant Prof and is from Biology Dept} \\ \beta_0 + \beta_4 + \epsilon & \text{if } i^{th} \text{person is ranked as Assistant Prof and is from Business Dept} \\ \beta_0 + \beta_1 + \beta_3 + \epsilon & \text{if } i^{th} \text{person is ranked as Associate Prof and is from Biology dept} \\ \beta_0 + \beta_1 + \beta_4 + \epsilon & \text{if } i^{th} \text{person is ranked as Associate Prof and is from Business dept} \\ \beta_0 + \beta_2 + \beta_3 + \epsilon & \text{if } i^{th} \text{person is ranked as Full Prof and is from Biology dept} \\ \beta_0 + \beta_2 + \beta_4 + \epsilon & \text{if } i^{th} \text{person is ranked as Full Prof and is from Business dept} \\ \beta_0 + \epsilon & \text{if } i^{th} \text{person is ranked as Assistant Prof and is from Family Studies dept} \end{cases}$$

L2.R

Take away messages

- Statistics:
  - Interactions: x1:x2 or (x1+x2)^2 or x1*x2
  - Dummy coding: the number of variable = the number of category -1
  - Interpretation of coefficients
- Code:
  - lm(y ~ x1+x2+(x1+x2)^2)
  - lm(y ~ factor(x1))

Thank you

- Questions OR Comments?

- Slack channel: section2-course-documents
- Email: qing.li2@uclagary.ca