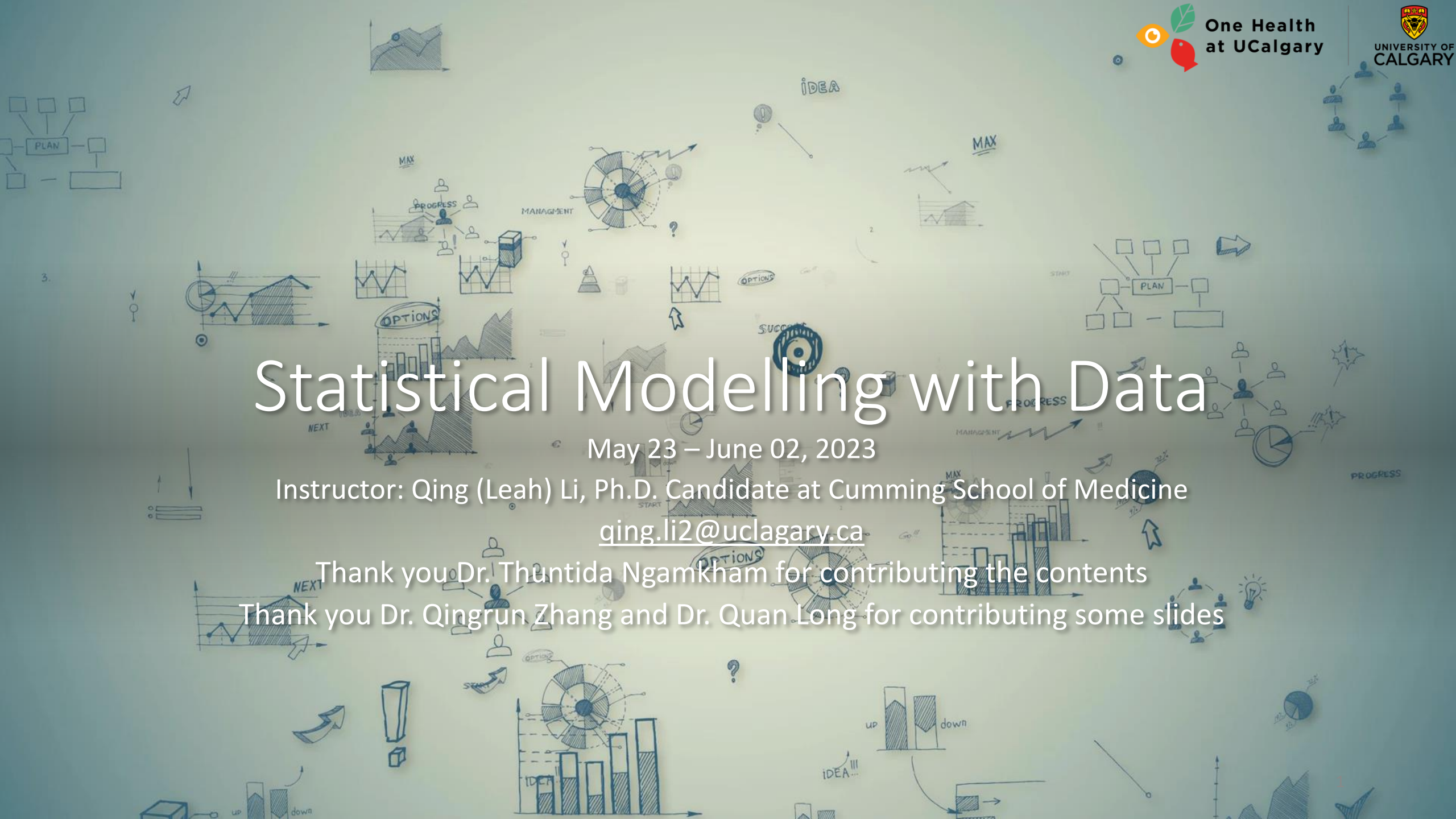


Statistical Modelling with Data

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

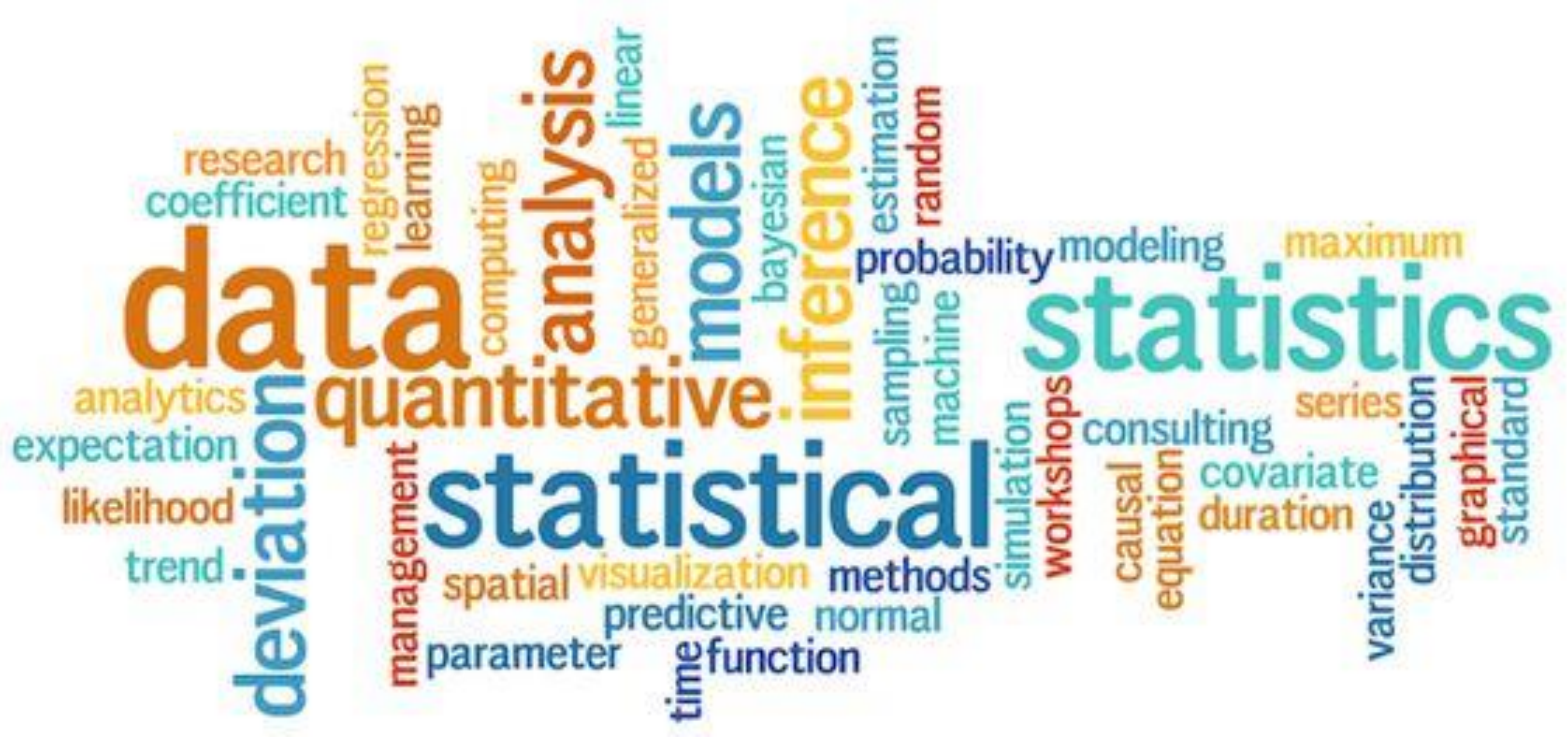
1. Model the multiple linear relationships between a response variable (Y) and all explanatory variables (both categorical and numerical variables) with interaction terms. Interpret model parameter estimates, construct confidence intervals for regression coefficients, evaluate model fits, and visualize 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.

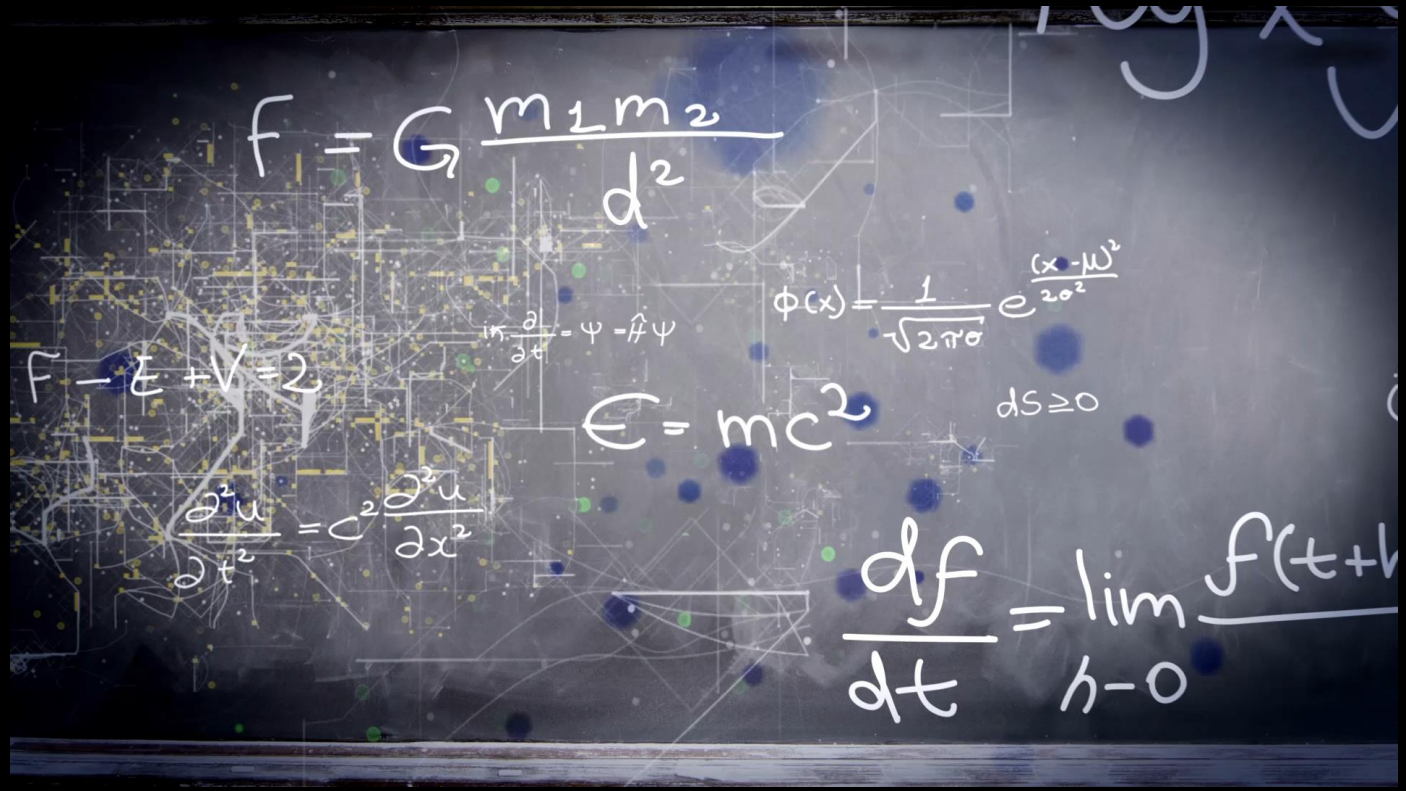
Statistical Modelling with Data

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

Motivations of your statistical modelling



Lecture 1: Multiple Linear Regression First-order models with quantitative independent variables



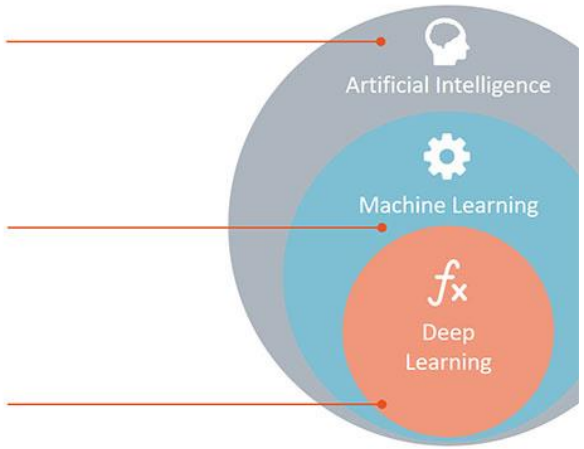
What is a statistical model?

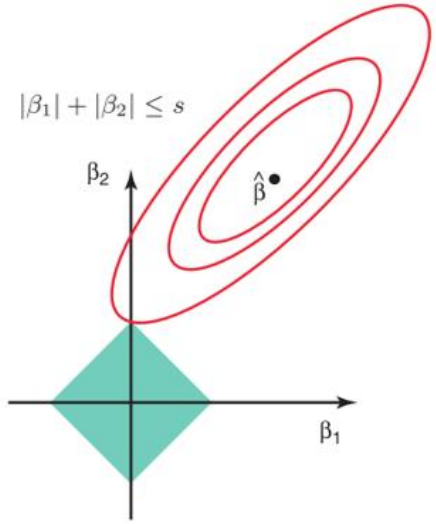
- A statistical model describes how a dependent variable (Y) is thought to have been generated.
- Parameter estimation is the process of computing a model's parameter values from measured data.

Artificial Intelligence
Any technique which enables computers to mimic human behavior.

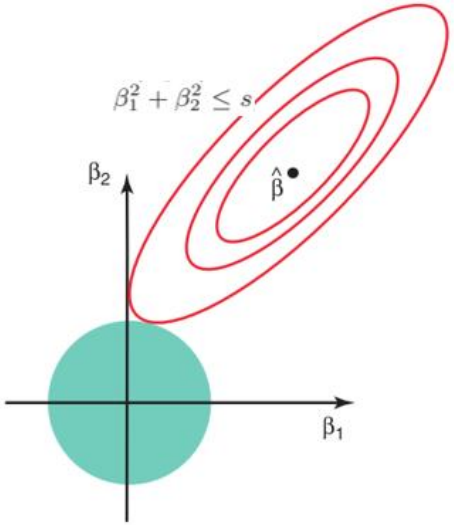
Machine Learning
Subset of AI techniques which use statistical methods to enable machines to improve with experiences.

Deep Learning
Subset of ML which make the computation of multi-layer neural networks feasible.



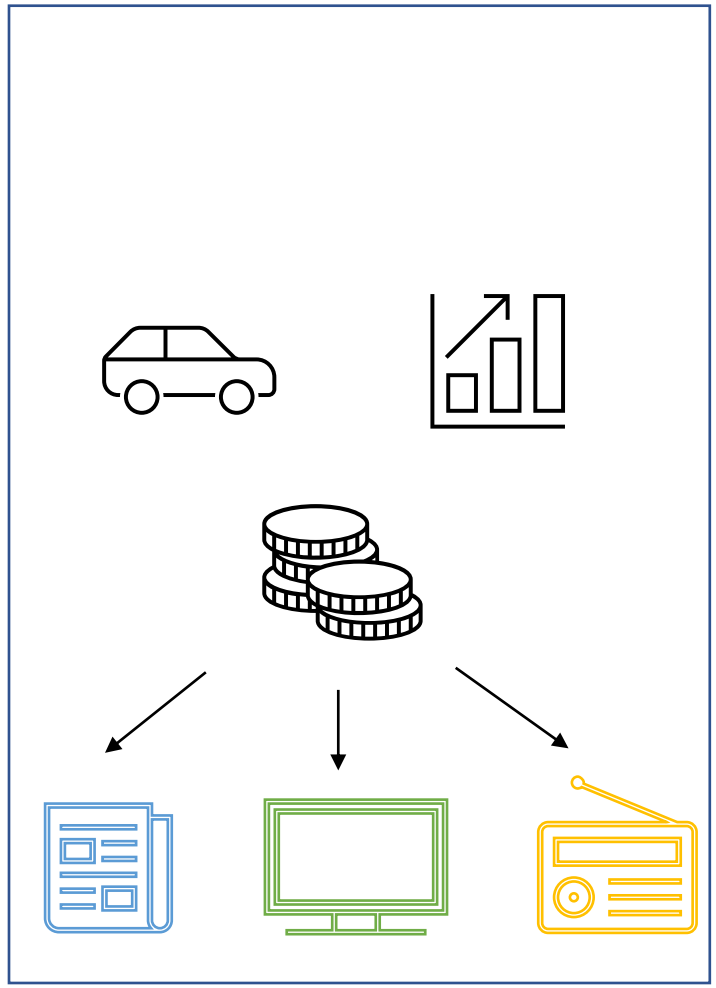


Lasso Regression



Ridge Regression

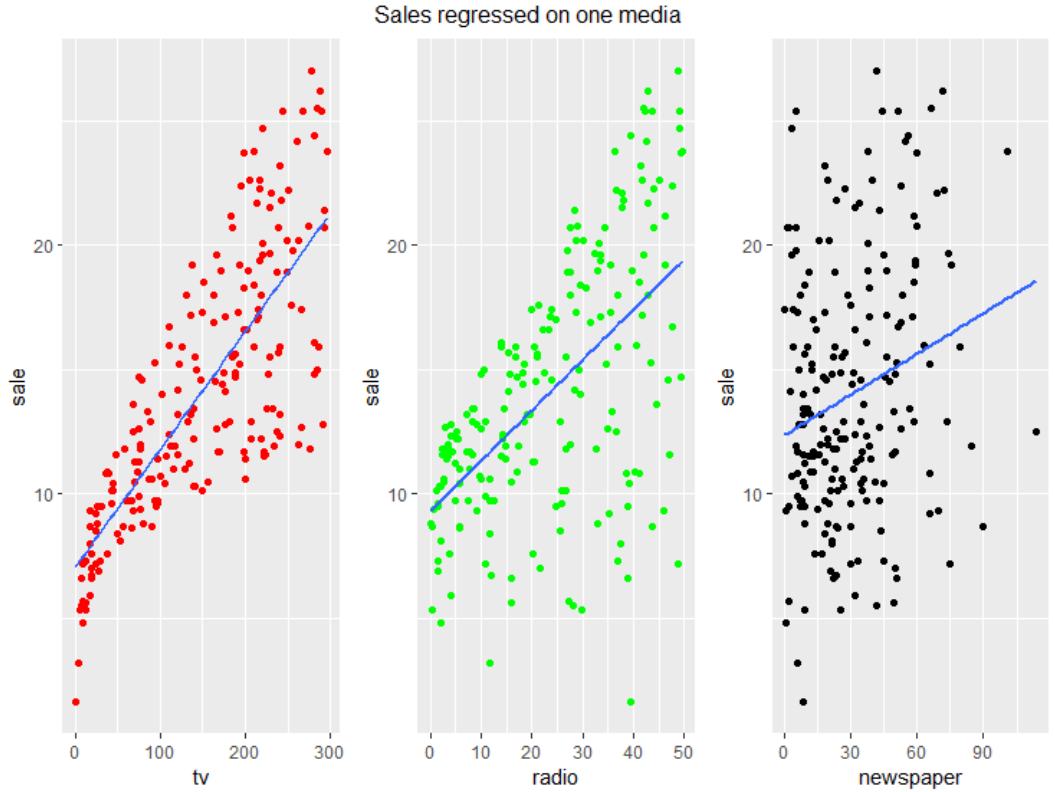
Simple linear model



Sales (in thousands of units)
Budgets(in thousands of dollars)

No	tv	radio	newspaper	sale
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75	7.2
7	57.5	32.8	23.5	11.8
8	120.2	19.6	11.6	13.2
9	8.6	2.1	1	4.8
10	199.8	2.6	21.2	10.6

Which factor affects sales most?



It seems advertainment budget through TV has a great impact on sales

sales ~ Newspaper advertisement budget

sales ~ TV advertisement budget

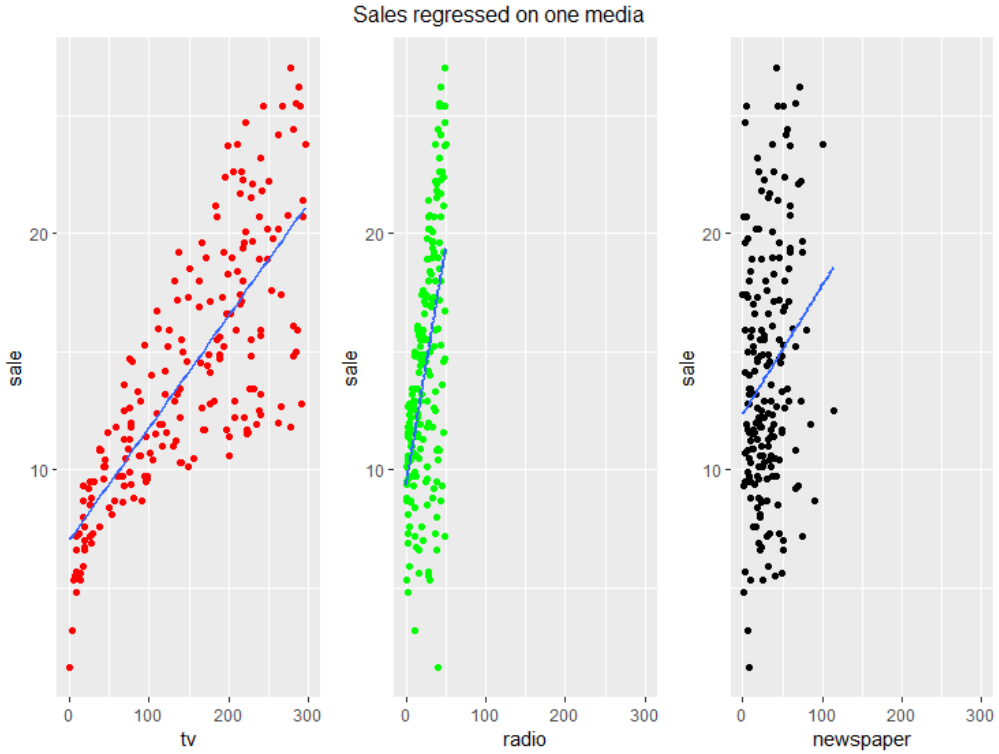
sales ~ Radio advertisement budget

Simple linear model

```
> reg1<-lm(sale~tv, data=Advertising)
> coefficients(reg1)
(Intercept)          tv
 7.03259355  0.04753664
>
> reg2<-lm(sale~radio, data=Advertising)
> coefficients(reg2)
(Intercept)       radio
  9.3116381   0.2024958
>
> reg3<-lm(sale~newspaper, data=Advertising)
> coefficients(reg3)
(Intercept)   newspaper
 12.3514071   0.0546931
```

```
summary(lm(sale~radio,data=Advertising))
```

```
##
## Call:
## lm(formula = sale ~ radio, data = Advertising)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7305  -2.1324   0.7707   2.7775   8.1810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.31164    0.56290  16.542   <2e-16 ***
## radio        0.20250    0.02041   9.921   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.275 on 198 degrees of freedom
## Multiple R-squared:  0.332,  Adjusted R-squared:  0.3287
## F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```



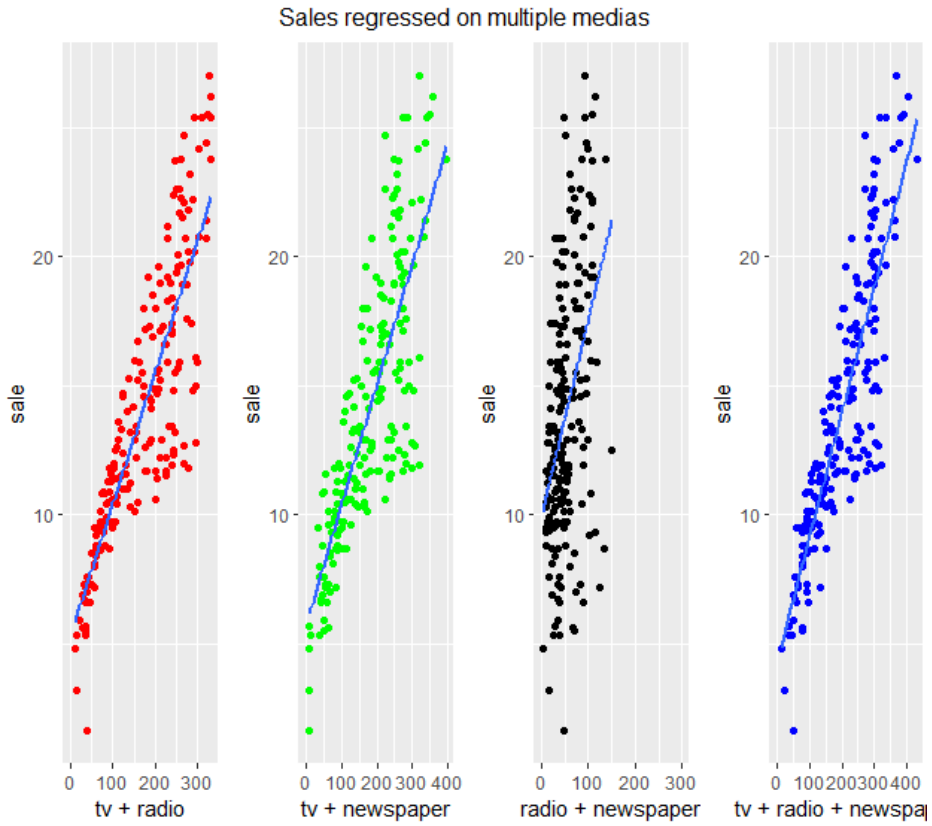
$$\hat{Sale} = 7.032594 + 0.047537tv$$

$$\hat{Sale} = 9.31164 + 0.20250radio$$

$$\hat{Sale} = 12.35141 + 0.05469newspaper$$

It's so easy! Wait a bit! Can TV and radio jointly affect sales?

Multiple linear model




```
> reg1<-lm(sale~tv+radio, data=Advertising)
> coefficients(reg1)
(Intercept)          tv       radio
 2.92109991  0.04575482  0.18799423
>
> reg2<-lm(sale~tv+newspaper, data=Advertising)
> coefficients(reg2)
(Intercept)          tv   newspaper
 5.77494797  0.04690121  0.04421942
>
> reg3<-lm(sale~radio+newspaper, data=Advertising)
> coefficients(reg3)
(Intercept)       radio   newspaper
9.188920459 0.199044594 0.006644175
>
> reg4<-lm(sale~tv+radio+newspaper, data=Advertising)
> coefficients(reg4)
 (Intercept)           tv        radio    newspaper
 2.938889369  0.045764645  0.188530017 -0.001037493
```

R codes:

lm() : "linear model" is used to create a simple or multiple regression model.

coefficients(): is used to extract model coefficients from a simple or multiple regression model.

The estimated model is $\hat{Sale} = 2.939 + 0.046tv + 0.189radio - 0.001newspaper$

We interpret these results as following:

$\hat{\beta}_1 = 0.046$ means that for a given amount of radio and newspaper advertising, spending additional \$1,000 on TV advertising leads to an *increase* in sales by approximately 46 units.

$\hat{\beta}_2 = 0.189$ means that for a given amount of TV and newspaper advertising, spending additional \$1,000 on radio advertising leads to an *increase* in sales by approximately 189 units.

$\hat{\beta}_3 = -0.001$ means that for a given amount of TV and radio advertising, spending additional \$1,000 on newspapers advertising leads to a *decrease* in sales by approximately 1 unit !!!!

Radio advertainment budget has the largest impact on sales

General linear model

- Statistics:
  - 1. continuous dependent variable ~ continuous independent variables (one/multiple)
  - 2. Pay attention to variables unit and scales
- R code:
  - lm(y~x)
  - Coefficients(lm(y~x))
  - summary (lm(y~x))
  - ggplot: https://ggplot2.tidyverse.org/
  - ggplot() : is used to construct the initial scatter plot.
  - geom_point(): the point geom is used to create scatterplots.
  - geom_smooth() : aids the eye in seeing patterns in the presence of overplotting.

A little bit mathematics

- General linear regression model, also called first-order model as it contains variables in first order format (in contrast to second order, third order, …)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$$

$$\begin{aligned}
where & \\
Y &= \text{ the dependent variable} \\
X_1, X_2, ..., X_p &= \text{ the independent variables, predictors} \\
E(Y) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \text{ the deterministic portion of the model} \\
\beta_i &= \text{regression coefficients}, i = 1, ..., p
\end{aligned}$$

From the Advertising example, Instead of fitting a separate simple linear regression model for each predictor, a better approach is to extend to the multiple linear regression model so that it can directly accommodate multiple predictors.

$$Sales = \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 newspapers + \epsilon$$

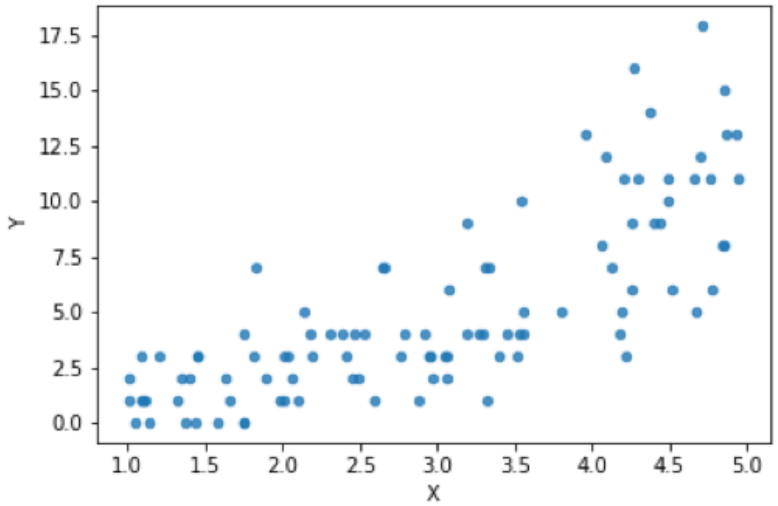
General linear model

- Simple linear regression is used to model the relationship between two continuous variables. Often, the objective is to predict the value of an output variable (or response) based on the value of an input (or predictor) variable.
- Multiple linear model is a regression model involving two or more independent variables.
- Simple linear regression and multiple linear regression model are general linear model. The term "general" linear model (GLM) usually refers to conventional linear regression models for a continuous response variable given continuous and/or categorical predictors.

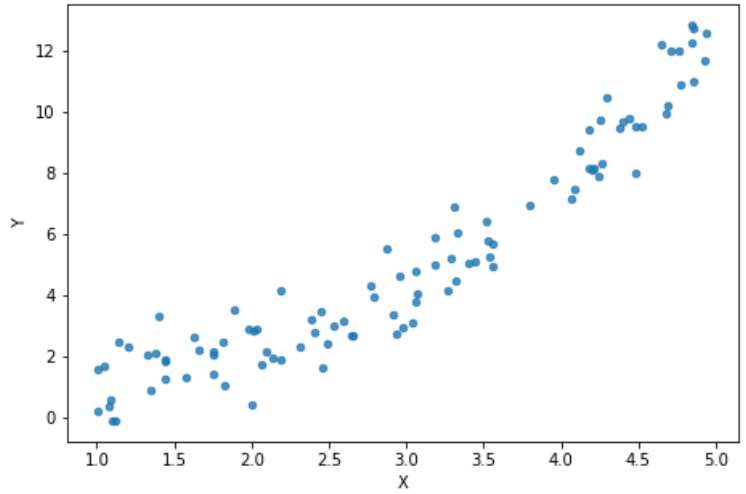
Limitations of the general linear models

- Linear models assume that
  - Y (generally) follows normal distribution
  - And the Xs impact Y in a linear manner
- This may be wrong if
  - Y is category (yes or no)
  - Y is a proportion of something
  - …
- So we need something more general

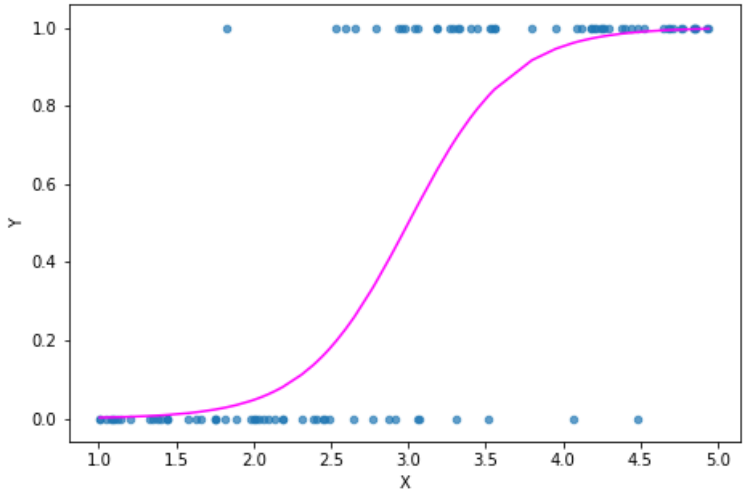
Limitations of the general linear models



$$g(y_i) = \beta_0 + \beta_1 X_1 + \varepsilon,$$
$$g(y_i) = y_i$$



$$g(y_i) = \beta_0 + \beta_1 X_1 + \varepsilon,$$
$$g(y_i) = \ln(y_i)$$

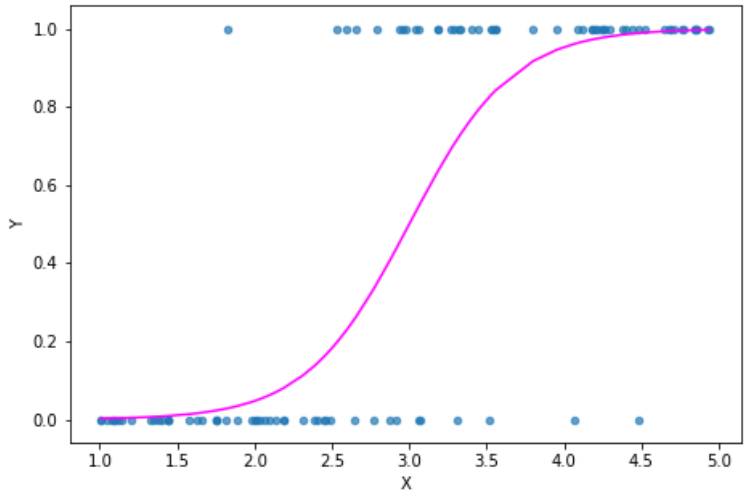


$$g(y_i) = \beta_0 + \beta_1 X_1 + \varepsilon,\ E(y_i) = p$$
$$g(y_i) = logit(p) = \ln\left(\frac{p}{1-p}\right)$$

Generalized linear model

- The response, Y, follows something not normal, precisely, an exponential family.
- Again, like the case of linear model, a set of Xs forming a linear predictor $\Sigma\beta_i X_i$
- Linking function: The relationship between the linear predictor and the mean of the distribution (that Y supposed to follow). For instance, logistic regression.
- "glm": for generalized linear regression

$$g(p) = \ln\left(\frac{p}{1-p}\right)$$



$$g(y_i) = \beta_0 + \beta_1 X_1 + \varepsilon, E(y_i) = p$$

$$g(y_i) = logit(p) = \ln\left(\frac{p}{1-p}\right)$$

An example for generalized linear model

- hours=c(0.50,0.75,1.00,1.25,1.50,1.75,1.75,2.00,2.25,2.50,2.75,3.00, 3.25,3.50,4.00,4.25,4.50,4.75,5.00,5.50)
- pass=c(0,0,0,0,0,0,1,0,1,0,1,0,1,0,1,1,1,1,1,1)
- glm_pass=glm(pass~hours, family=binomial(link='logit'))

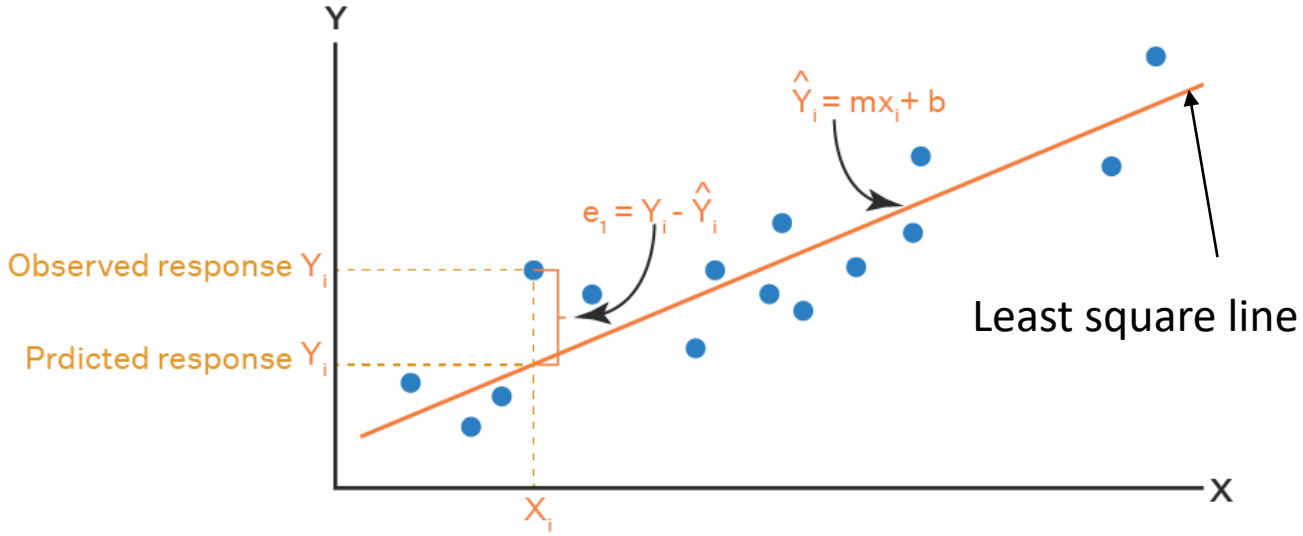
```
> coefficients(glm_pass)
(Intercept)       hours
  -4.077713    1.504645
```

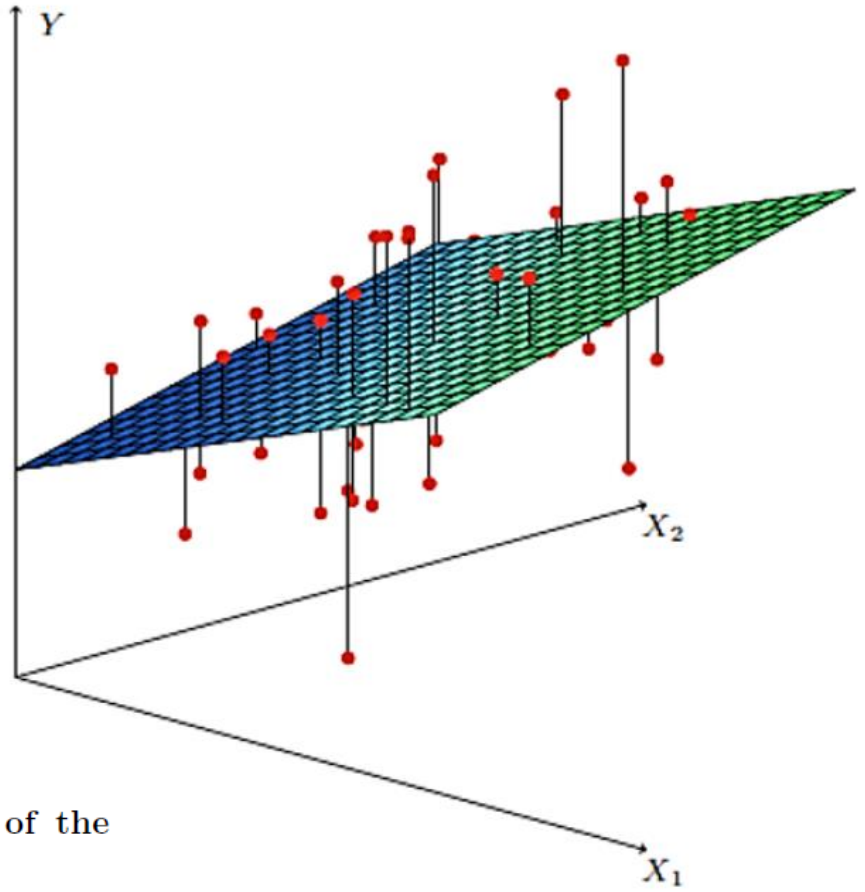
P(passing the exam) = 1/(1+exp(-(1.50 *hours -4.08)))

Least square method

The least squares method is a statistical procedure to find the best fit for a set of data points by minimizing the sum of the offsets or residuals of points from the plotted curve.

Parameter estimation is the process of computing a model's parameter values from measured data.





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:**

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

Confidence interval

- Is the estimated value equals to actual value?
- Can we find an interval where the actual value is in?
- Confidence interval (CI) refers to a range of values within which statisticians believe the actual value of a certain population parameter lies.

$$(1-\alpha)\%\ Confidence\ Interval\ for\ parameter\ \beta_i\ is\ \hat{\beta}_i \pm t_{\alpha/2} S_{\hat{\beta}_i}$$

- 95% confidence interval for a given parameter: if we conduct 100 tests, in 95 tests, our calculated interval contain the actual value. However, in the rest 5 tests, our calculated intervals do not contain the actual value.
- The confidence interval is based on the observations from a test, and hence differs from test to test.
- We cannot say “with probability (1 – α), the parameter μ lies in the confidence interval.” Because once the CI is calculated, the actual value is either in the CI or not.
- In real life, statisticians often have limited number of tests. With 95% CI, analysts think they have a good chance of including the real number.
- The figure on the right shows 50 realizations of a confidence interval [can be long or short] for a given population mean μ. 95% confidence interval for mu requires 100 realizations, 95 realizations CI contains true mu and 5 not.

Confidence interval

```
reg1<-lm(sale~tv+radio+newspaper, data=Advertising)
confint(reg1) # a 95% confidence interval for coefficients
```

```
##                 2.5 %     97.5 %
## (Intercept)  2.32376228 3.55401646
## tv           0.04301371 0.04851558
## radio        0.17154745 0.20551259
## newspaper   -0.01261595 0.01054097
```

```
confint(reg1, level = 0.99) # a 99% confidence interval for coefficients
```

```
##                 0.5 %     99.5 %
## (Intercept)  2.12757072 3.75020802
## tv           0.04213632 0.04939297
## radio        0.16613095 0.21092909
## newspaper   -0.01630884 0.01423386
```

In class Practice Problem 1

How do real estate agents decide on the asking price for a newly listed condominium?

A computer data base in a small community contains the *listed selling price* (in thousand of dollars), the *amount of living area* (in hundreds of square metres), and *the number of floors*, *bedrooms*, and *bathroom* are recorded for 15 randomly selected condos currently on the market. The data file is provided in **condominium.csv**.

a) Use R to fit a model explaining selling price (Y) with all the available explanatory variables.

b) Construct a 95% confidence interval for regression coefficients.

L1.R

Coffee break

Come back at 14:40pm?

Models Evaluations

1. Is this multiple regression model any good at all? Is at least one of the predictors useful in predicting the response?

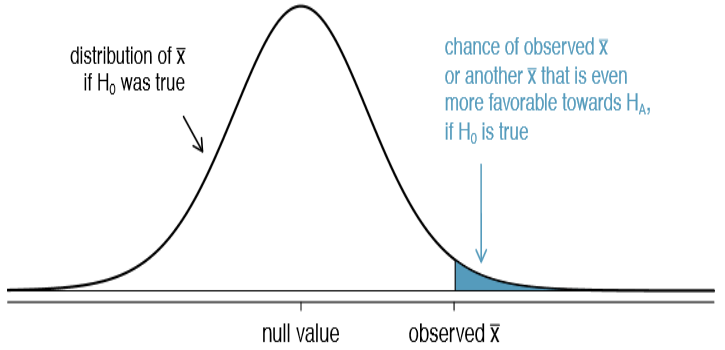
2. Do all the predictors help to explain Y , or is only a subset of the predictors useful?

Evaluating Overall Model Utility

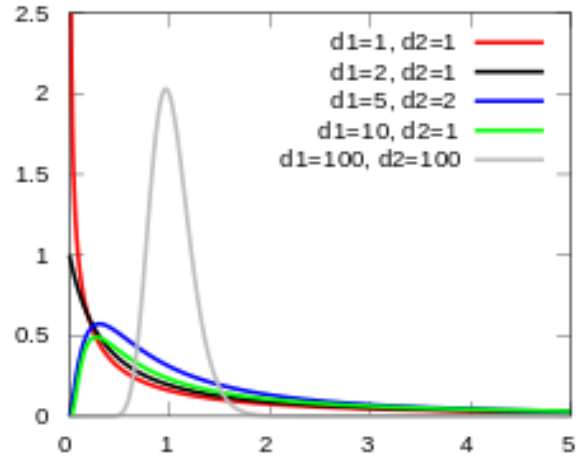
1. Is this multiple regression model any good at all? Is at least one of the predictors useful in predicting the response?

- The first of these hypotheses is an overall F-test or a global F test which tells us if the multiple regression model is useful.

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$
$$H_a : \text{at least one } \beta_i \text{ is not zero } (i = 1, 2, ..., p)$$



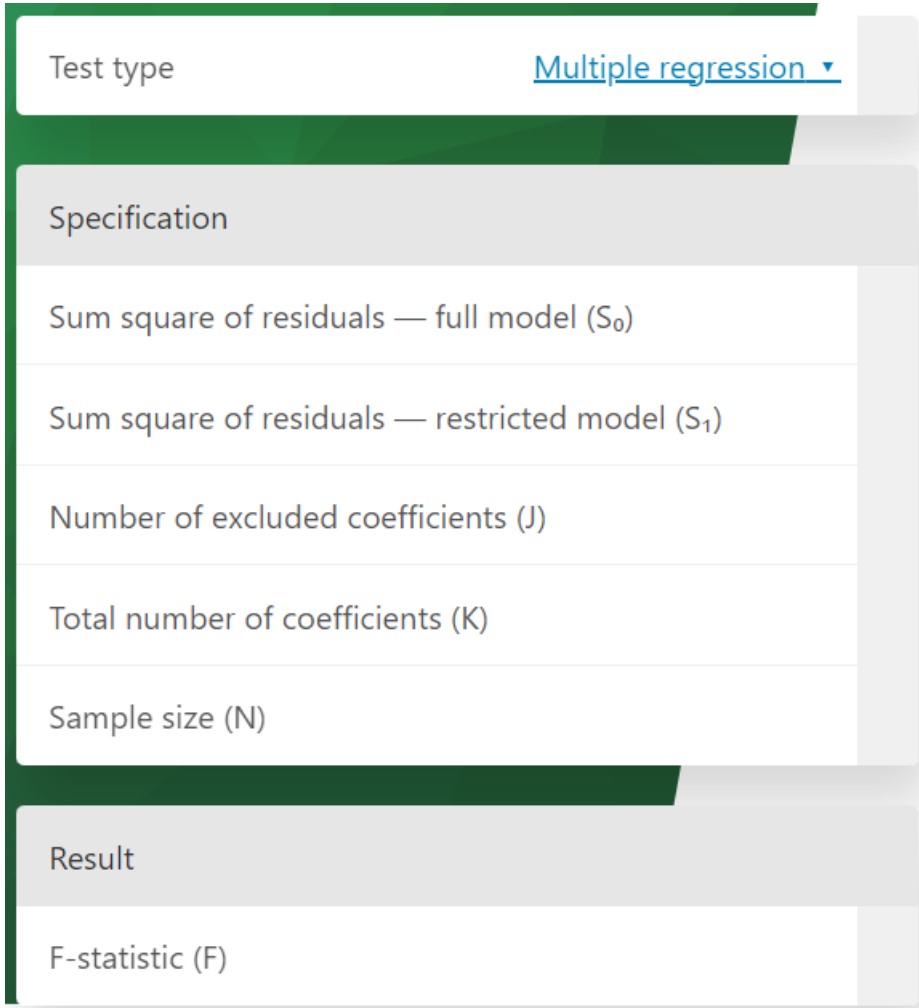
- An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis.
- A test statistic is a statistic (a quantity derived from the sample) used in statistical hypothesis testing.
- F test statistic: $\mathrm{F} = \frac{X_1/\vartheta_1}{X_2/\vartheta_2}, X_1 \sim \chi^2, df1, X_2 \sim \chi^2, df2$



Evaluating Overall Model Utility

F critical values

Degrees of freedom in the denominator	p	Degrees of freedom in the numerator: 1	2	3	4	5	6	7	8	9
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
	.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
	.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
	.001	405284	500000	540379	562500	576405	585937	592873	598144	602284
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39
3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86
4	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47
5	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24



https://www.omnicalculator.com/statistics/f-statistic

Evaluating Overall Model Utility

- ANOVA test to compare several (more than two) groups' means.
  - One-way: one independent variable
  - Two-way: two independent variable
- Comparison of two models

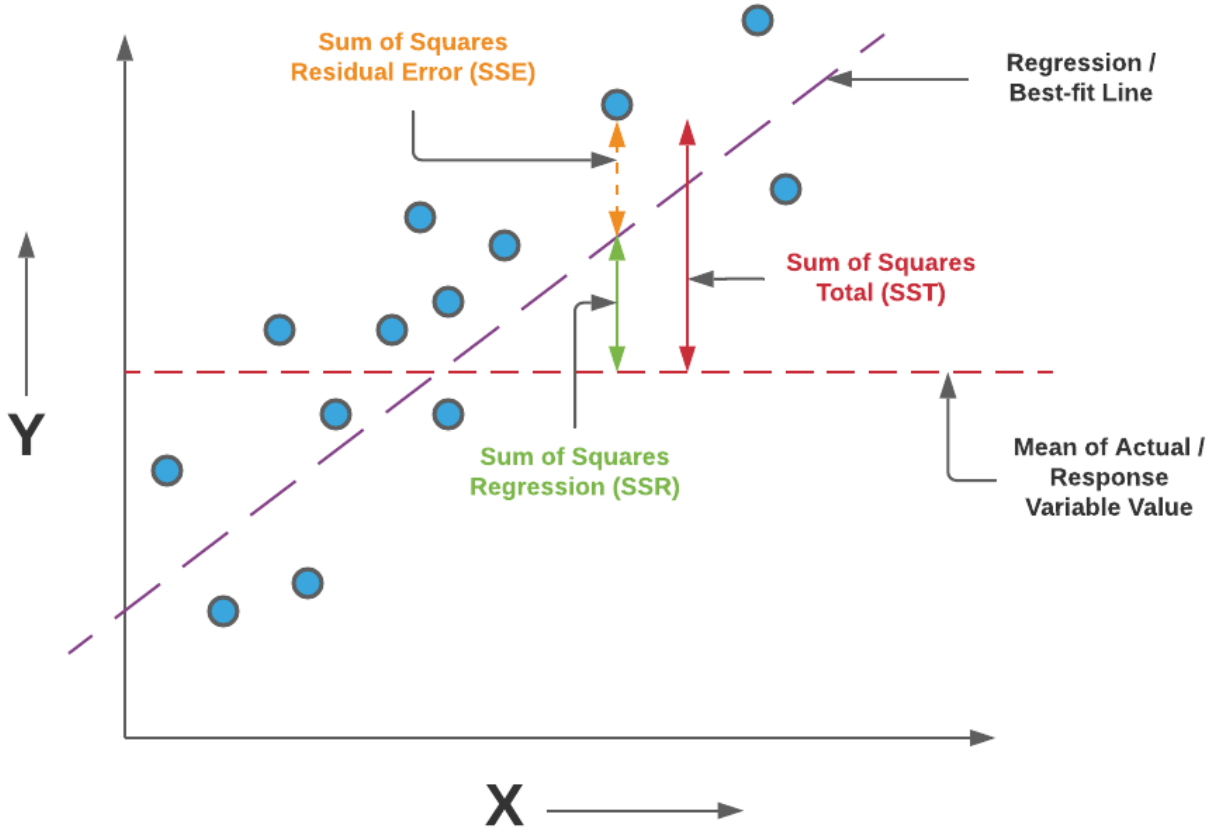
F_{df2}^{df1} or $F(df1, df2)$.

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		

$$F_{cal} = \frac{MSR}{MSE} = \frac{\frac{SSR}{p}}{\frac{SSE}{(n-p-1)}}$$

Evaluating Overall Model Utility



Sum of squares for error or residual $= SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

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

Sum of squares for regression $= SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

Total corrected sum of squares of the Y's $= SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$

n = the sample size

p = the number of predictors or the number of regression coefficients

$$SST = SSR + SSE$$

$$F_{cal} = \frac{MSR}{MSE} = \frac{\frac{SSR}{p}}{\frac{SSE}{(n-p-1)}}$$

The larger sample variance always goes in the numerator to make the right-tailed test, and the right-tailed tests are always easy to calculate.

Evaluating Overall Model Utility

```
> reg1<-lm(sale~tv+radio+newspaper, data=Advertising) # (Full) model with all variables
> reg2<-lm(sale~1, data=Advertising) # Model with only intercept
> anova(reg2,reg1)
Analysis of Variance Table

Model 1: sale ~ 1
Model 2: sale ~ tv + radio + newspaper
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    199 5417.1
2    196  556.8  3    4860.3 570.27 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA table for Advertising data example

Source of Variation	DF	Sum of Squares	Mean Square	F-Statistic
Regression	3	4860.3	1620.1	570.295
Residual	196	556.8	2.84081	
Total	199	5417.1		

Evaluating effect of subset of predictors

2. Do all the predictors help to explain Y , or is only a subset of the predictors useful?

- T-test, identify the effect of each individual predictor

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

$$T = \frac{Sample\ statistic - assume\ population\ mean}{estimated\ standard\ error\ of\ statistic}$$

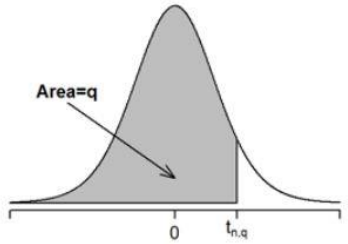
$$t_{cal} = \frac{\widehat{\beta}_i - \beta_i}{SE(\widehat{\beta}_i)} \text{ which has } df = n - p \text{ degree of freedom}$$

$$SE_{\widehat{\beta}} = \frac{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

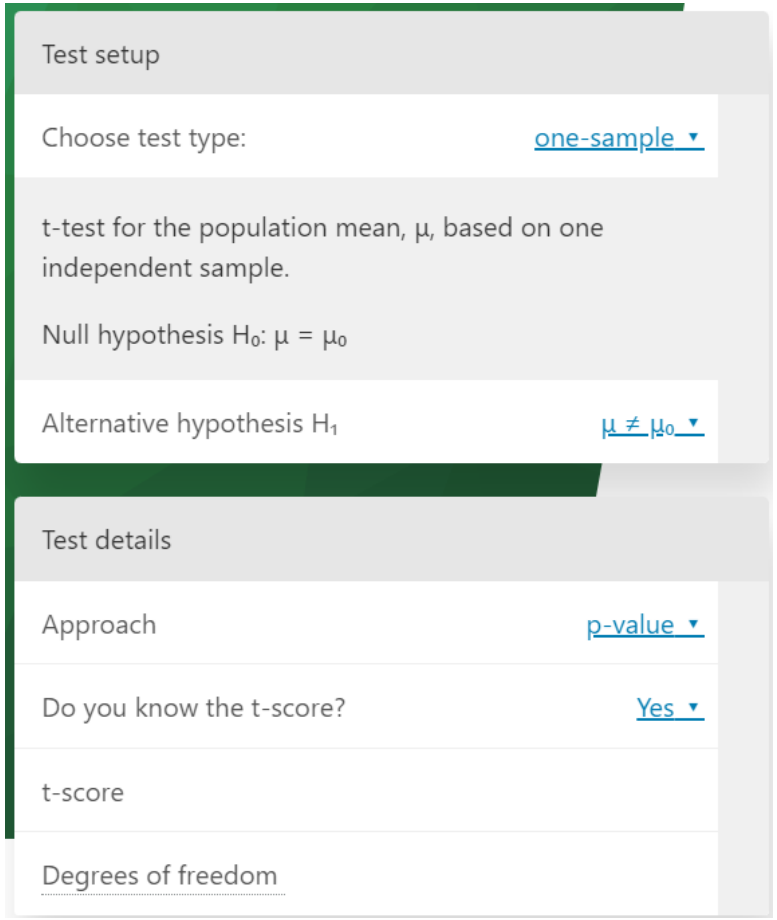
Evaluating effect of subset of predictors

Quartiles of the t Distribution

The table gives the value if $t_{n;q}$ - the qth quantile of the t distribution for n degrees of freedom



	q = 0.6	0.75	0.9	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995
n = 1	0.3249	1.0000	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.2887	0.8165	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.2767	0.7649	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.2707	0.7407	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.2672	0.7267	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.2648	0.7176	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.2632	0.7111	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.2619	0.7064	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.2610	0.7027	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.2602	0.6998	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.2596	0.6974	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.2590	0.6955	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.2586	0.6938	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.2582	0.6924	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140



https://www.omnicalculator.com/statistics/t-test

Evaluating effect of subset of predictors

2. Do all the predictors help to explain Y , or is only a subset of the predictors useful?

- Partial F-test, identify the usefulness of a subset of predictors

$$F_{cal} = \frac{\frac{SSE_{\text{reduced model}} - SSE_{\text{full model}}}{df_{reduced} - df_{full}}}{\frac{SSE_{\text{full model}}}{df_{full}}}$$

Full Model to be the model with the whole set of predictors

Reduced Model to be the model with the whole set of predictors less the subset to be tested.

For example, if we want to test X1 given X2 and X3 are in the model, then the Full Model has the predictors X1 , X2 and X3, and the Reduced Model has the predictors X2 and X3. This will test the effect of X1 in the full model with all 3 predictors.$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$. The hypotheses are:

$$H_0 : \beta_1 = 0 \text{ in the model } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$
$$H_a : \beta_1 \neq 0 \text{ in the model } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

In general,to test that a particular subset of q of the coefficients are zero, the hypotheses are

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = ... = \beta_p = 0$$
$$H_a : \text{at least one } \beta_i \neq 0$$

This can be achieved using an F-test. Let SSE(Full model) be the residual sum of squares under the full model and SSE(Reduced model) be the residual sum of squares under the reduced model. Then the F-statistic is

Evaluating effect of subset of predictors

```
> full<-lm(sale~tv+radio+newspaper, data=Advertising)
> reduced<-lm(sale~tv+radio, data=Advertising) # dropping a newspaper variable
> anova(reduced,full) # test if Ho: newspaper = 0
Analysis of Variance Table

Model 1: sale ~ tv + radio
Model 2: sale ~ tv + radio + newspaper
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    197 556.91
2    196 556.83  1  0.088717 0.0312 0.8599
```

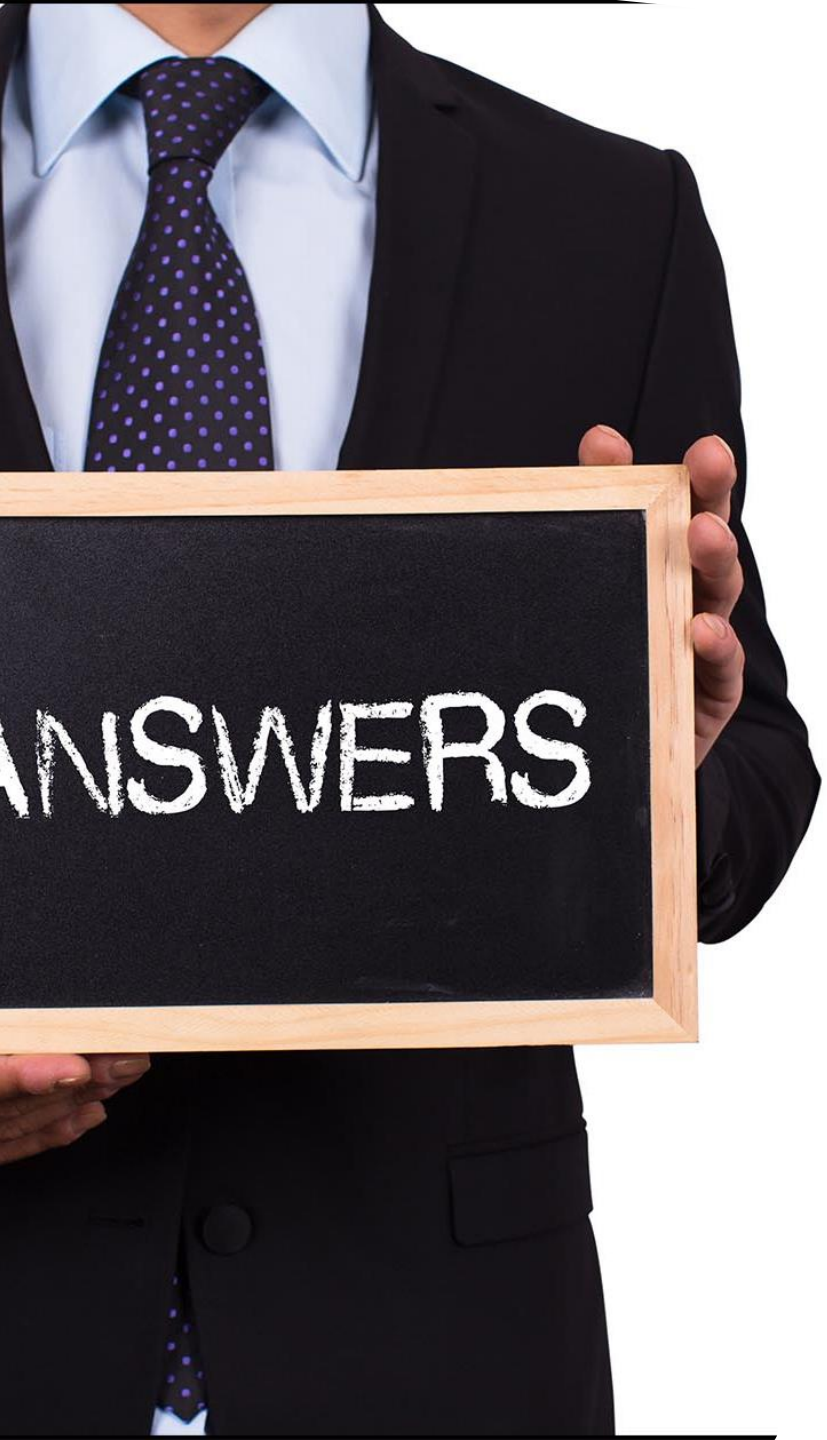
The ANOVA table for Advertising models

Source of Variation	DF	Sum of Squares	Mean Square	F-Statistic
Regression	1	0.088717	0.088717	0.0312
Residual	196	556.83	2.84081	
Total	197	556.91		

- With df=1,196 (p-value=0.8599 > α = 0.05), indicating that we should clearly not to reject the null hypothesis which mean that we can definitely drop the variable newspaper off the model.
- At this point, from the initial estimated regression model is *Sales* = 2.939 + 0.046*tv* + 0.189*radio* − 0.001*newspaper*
- After checking individual coefficients test, the final regression model is *Sales* = 2.92110 + 0.04575*tv* + 0.18799*radio*

In class Practice Problem 2

- Use the condominium data (condominium.csv)
- Use the method of Partial F test to fit the model.
- How many possible fitted models would you suggest for predictive purposes?



In class Practice Problem 2

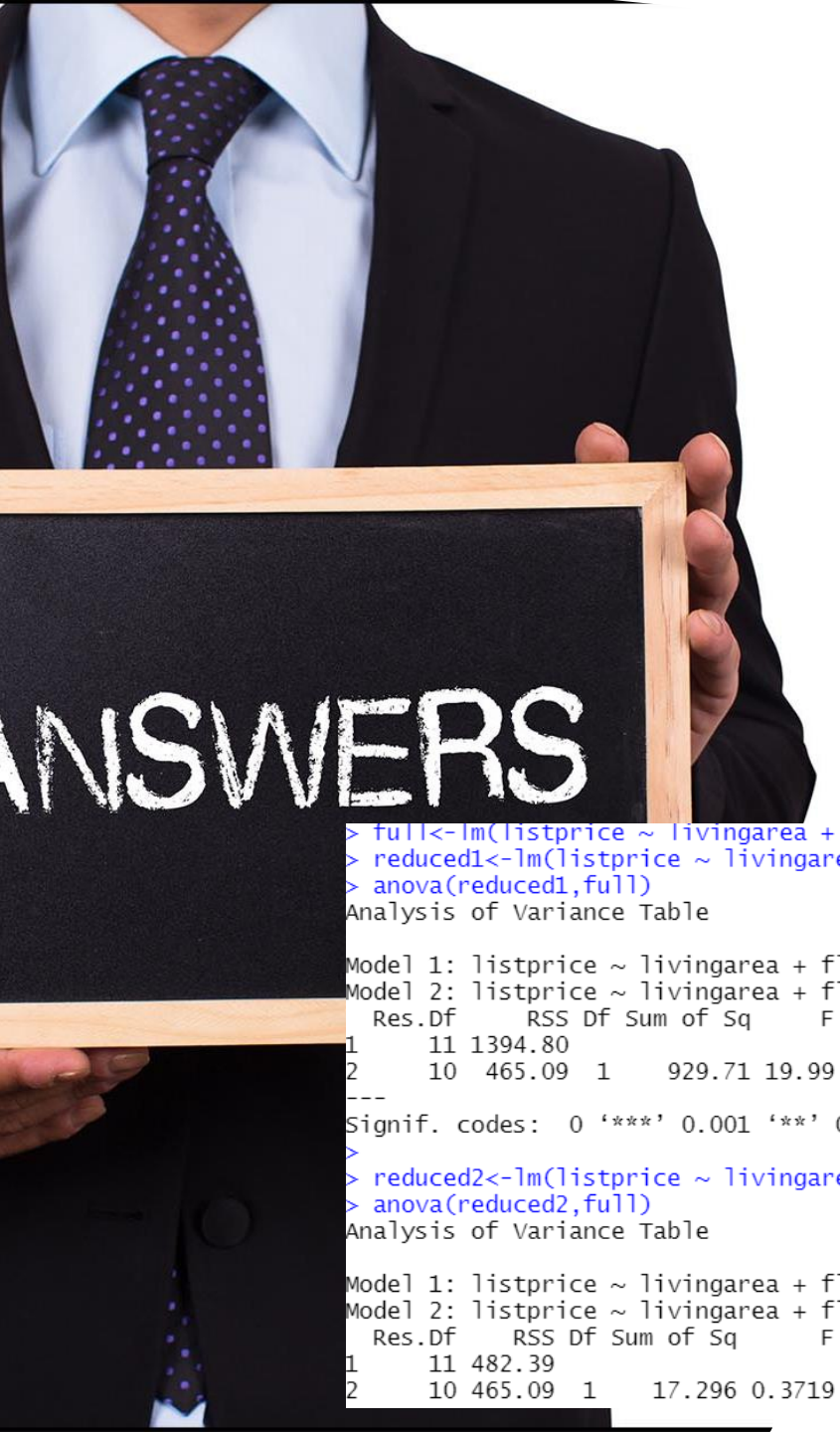
```
> model1 = lm(listprice ~ livingarea + floors + bedrooms + baths, data = condominium)
> summary(model1)

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

Residuals:
    Min      1Q  Median      3Q     Max
-12.617  -1.661   1.114   2.411  11.833

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   18.597      9.165   2.029   0.0699 .
livingarea    67.678      7.790   8.688 5.68e-06 ***
floors       -16.508      6.198  -2.664   0.0237 *
bedrooms      -2.730      4.477  -0.610   0.5556
baths         30.479      6.817   4.471   0.0012 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.82 on 10 degrees of freedom
Multiple R-squared:  0.9716,    Adjusted R-squared:  0.9603
F-statistic: 85.56 on 4 and 10 DF,  p-value: 1.08e-07
```



In class Practice Problem 2

Partial F-test

```
> full<-lm(listprice ~ livingarea + floors + bedrooms + baths, data=condominium)
> reduced1<-lm(listprice ~ livingarea + floors + bedrooms, data=condominium)
> anova(reduced1,full)
Analysis of Variance Table

Model 1: listprice ~ livingarea + floors + bedrooms
Model 2: listprice ~ livingarea + floors + bedrooms + baths
  Res.Df     RSS Df Sum of Sq     F   Pr(>F)
1     11 1394.80
2     10  465.09  1    929.71 19.99 0.001196 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> reduced2<-lm(listprice ~ livingarea + floors+ baths , data=condominium)
> anova(reduced2,full)
Analysis of Variance Table

Model 1: listprice ~ livingarea + floors + baths
Model 2: listprice ~ livingarea + floors + bedrooms + baths
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     11 482.39
2     10 465.09  1    17.296 0.3719 0.5556
```

>>Drop baths

>>Drop bedrooms

```
> reduced3<-lm(listprice ~ livingarea + baths , data=condominium)
> anova(reduced3,reduced2)
Analysis of Variance Table

Model 1: listprice ~ livingarea + baths
Model 2: listprice ~ livingarea + floors + baths
  Res.Df    RSS Df Sum of Sq    F  Pr(>F)
1     12 809.54
2     11 482.39  1    327.15 7.46 0.01953 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> reduced4<-lm(listprice ~ floors + baths , data=condominium)
> anova(reduced4,reduced2)
Analysis of Variance Table

Model 1: listprice ~ floors + baths
Model 2: listprice ~ livingarea + floors + baths
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     12 4968.2
2     11  482.4  1    4485.8 102.29 6.602e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> reduced5<-lm(listprice ~ livingarea + floors , data=condominium)
> anova(reduced5,reduced2)
Analysis of Variance Table

Model 1: listprice ~ livingarea + floors
Model 2: listprice ~ livingarea + floors + baths
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     12 1563.79
2     11  482.39  1    1081.4 24.659 0.0004249 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

>>Drop floors

>>Drop living area

>>Drop baths

Model goodness of fit

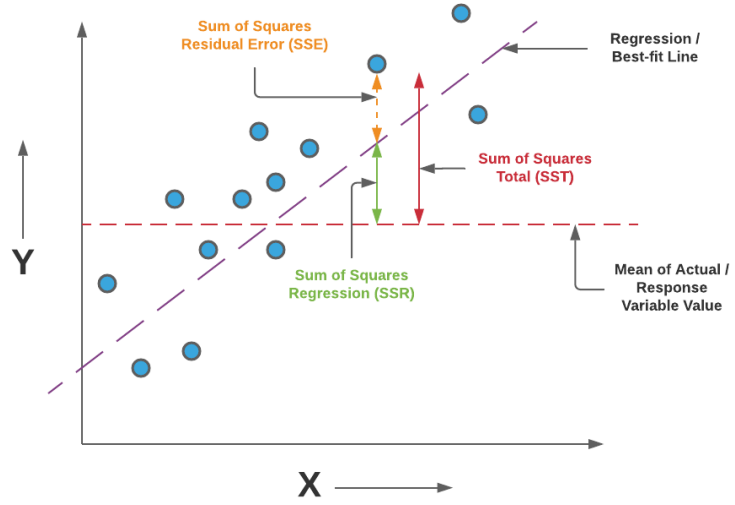
- How well does the regression model fit?? Two of the most common numerical measures of model fit are RSE(Residual Standard Error: *s*) and *R*2 (Coefficient of Determination), the fraction of variation explained. These quantities are computed and interpreted in the same fashion as for simple linear regression.

R^2(the Coefficient of Determination)

Recall that in simple linear regression, R^2 is the square of the correlation of the response and the variable. **In multiple regression**, it turns out that it equals to $Cor(Y, \hat{Y})^2$, the square of the correlation between the response and the fitted linear model, R^2 is the proportion of the total variation that is explained by the regression model of Y on $X_1, X_2, ...X_p$ that is,

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

An R^2 value close to 1 indicates that the model explains a large portion of the variance in the response variable. For example, if R^2 is 0.7982 for the model, then 79.82% of the variation of the response variable is explained by the model.



Model goodness of fit

It turns out that R^2 will always increase when more variables are added to the model, even if those variables are only weakly associated with the response. To compensate for this one can define **an adjusted coefficient of determination**, R^2_{adj}

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}}$$

```
> full<-lm(sale~tv+radio+newspaper, data=Advertising)
> reduced<-lm(sale~tv+radio, data=Advertising)
> summary(full)$r.squared
[1] 0.8972106
> summary(reduced)$r.squared
[1] 0.8971943
> summary(full)$adj.r.squared
[1] 0.8956373
> summary(reduced)$adj.r.squared
[1] 0.8961505
```

Model goodness of fit

One way to assess strength of fit is to consider how far off the model is for a typical case. That is, for some observations, the fitted value will be very close to the actual value, while for others it will not. The magnitude of a typical residual can give us a sense of generally how close our estimates are. Some of the residuals are positive, while others are negative.Thus, it makes more sense to compute the square root of the mean squared residual and to make this estimate unbiased, we have to divide the sum of the squared residuals by the degrees of freedom in the model. In general, RMSE or s is defined as

$$s = RMSE = \sqrt{\frac{1}{n-p-1} SSE} = \sqrt{MSE}$$

where

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

RMSE can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable. **Lower values of RMSE indicate better fit.**

```
> #MSE
> full<-lm(sale~tv+radio+newspaper, data=Advertising)
> reduced<-lm(sale~tv+radio, data=Advertising)
> sigma(full)
[1] 1.68551
> sigma(reduced)
[1] 1.681361
```

Model prediction

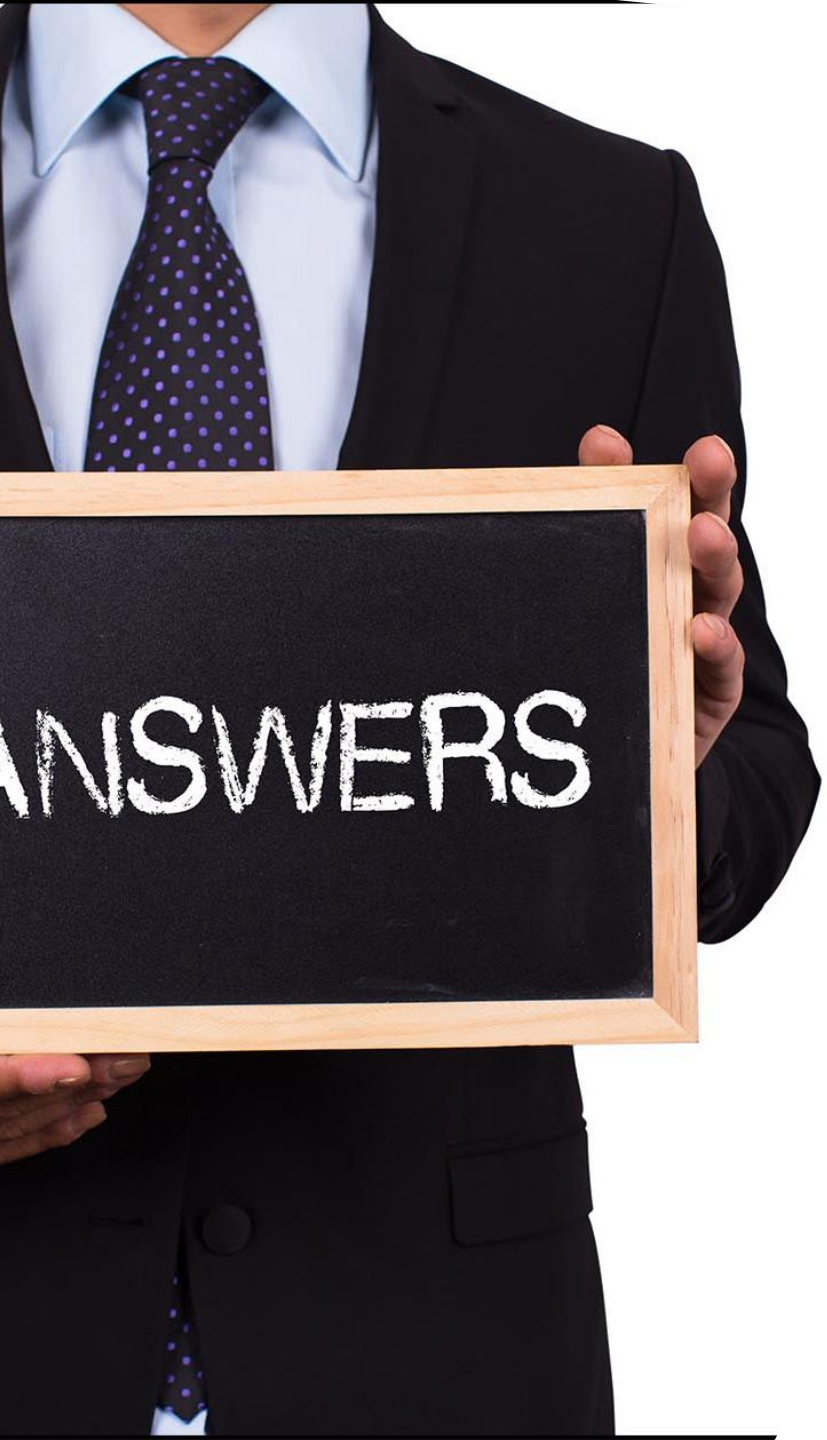
- Once we have fit the multiple regression model, it is straightforward to predict the response Y on the basis of a set of values for the predictors *X*1,*X*2, ...*Xp*. We usually use a **prediction interval** to predict the response y

```
> 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
```

- The 95% confidence interval of the sale with the given parameters is between 12.5042 (thousand units)and 19.1597(thousand units) when the TV and Radio advertising budgets are 200 thousand dollars and 20 thousand dollars, respectively.

In class Practice Problem 3

- Use the condominium data
- Use the method of Model Fit to calculate *R2adj* and RMSE for all possible models.
- Which model or set of models would you suggest for predictive purposes?
- Provide the 95% condominium list price prediction interval for your dream house.

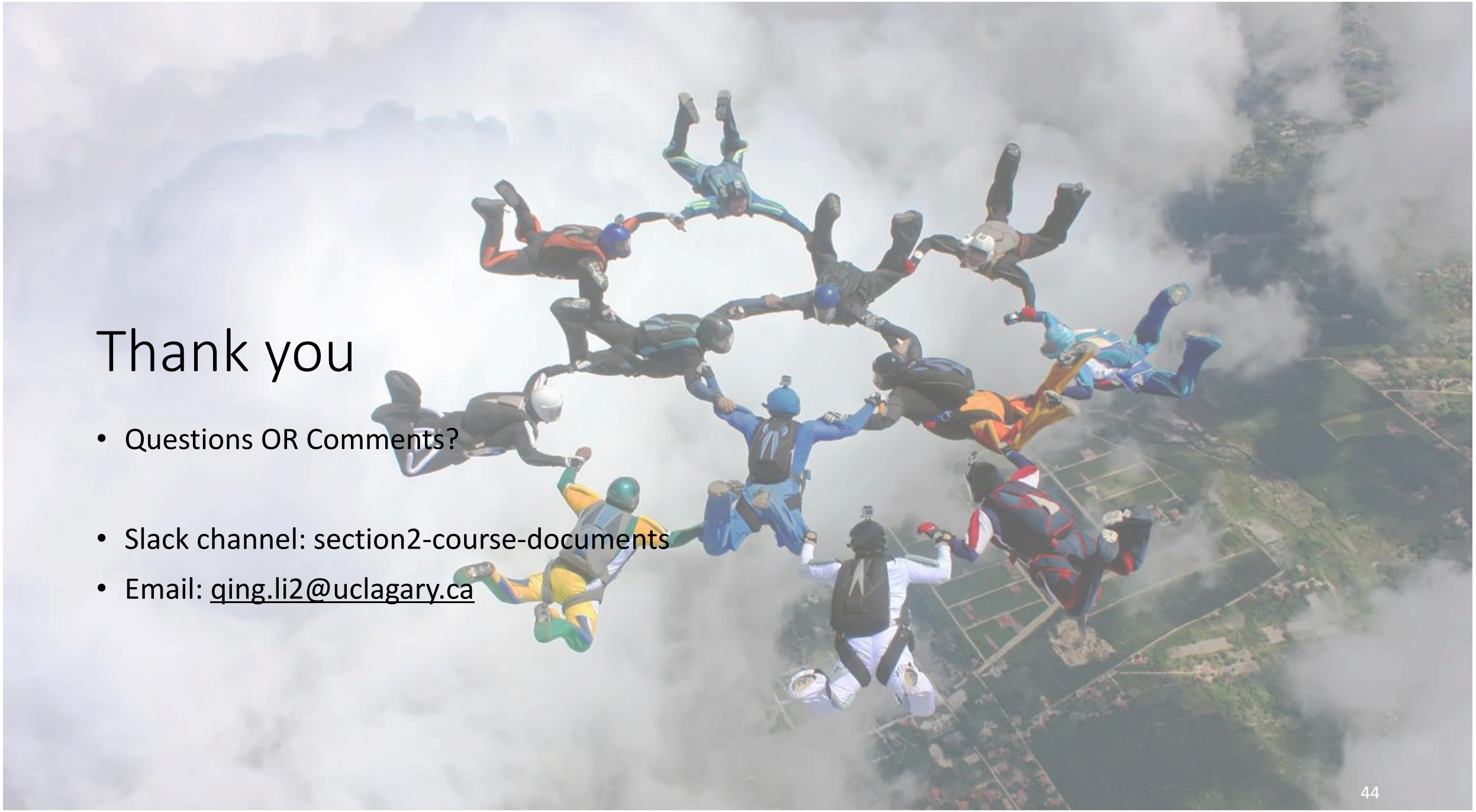


In class Practice Problem 3

```
> 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
```

Take away messages

- Statistics:
  - General linear model/Generalized linear model
  - Least square method to estimate parameters
  - Model utility (F-test, ANOVA)
  - Model goodness of fit: R^2, R^2_{adj}, MSE
  - Model prediction
- Code:
  - lm(); glm(); summary(); coefficients(); confint(); annova(); predict();



Thank you

- Questions OR Comments?

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