Exam 1 Review

Exam 1

Cathy Poliak, Ph.D.
cpoliak@central.uh.edu

Department of Mathematics
University of Houston

Exam structure.

- Thursday 10/6 at 11:30 am in GAR 201 during class.
- Approximately 8 questions
- 75 minutes
- May bring one-page notes front/back can be typed if wanted to be turned in with the test for bonus points. Only notes, formulas and R code no worked out examples.
- Bring your calculator.

Three problems will present you with a data example and ask you an array of modeling/interpretation questions about that data. (Short answer questions)

Other problems will just be a mix of single questions on general knowledge of the class material. Will be a mixture of multiple choice and short answer questions.

Topics Covered

- Types of statistical learning
- Simple linear regression
- Multiple linear regression
- Polynomial regression
- Best subsets
- Logistic Regression
- Test/Training data
- Confusion Matrix
- Linear Discriminant Analysis (LDA)

Type of Statistical Learning

In many data problems we are faced with one of two tasks:

- Prediction
- Inference

Are the following problems a) Prediction or b) Inference?

1. Explain what factors cause cancer $\rightarrow$? Inference
2. Forecast the weather $\rightarrow$? Prediction
3. Predict freshman's final college GPA $\rightarrow$? Prediction
4. Explain what factors affect college GPA $\rightarrow$? Infrence

Prediction Versus Inference

In prediction, our sole and primary goal is to predict well at all costs, no matter the interpretation of underlying mechanism.

Prediction Versus Inference

In prediction, our sole and primary goal is to predict well at all costs, no matter the interpretation of underlying mechanism.

In inference, our goal is to infer the relationship between variables and response, to estimate population parameters (μ, or β_0, β_1 etc). Interpretation is king for inference, most times at a cost of a worse prediction performance.

Response and predictor, or explanatory, variables.

Having mentioned prediction and inference tasks, we have to ask:

- What are we predicting? $\Longrightarrow$

Response and predictor, or explanatory, variables.

Having mentioned prediction and inference tasks, we have to ask:

- What are we predicting? $\Longrightarrow$ response variable
- With help of what are we predicting the response variable? $\Longrightarrow$

Response and predictor, or explanatory, variables.

Having mentioned prediction and inference tasks, we have to ask:

- What are we predicting? $\Longrightarrow$ response variable
- With help of what are we predicting the response variable? $\Longrightarrow$ predictor, or explanatory, variables
- In inference, we are inferring the relationships between...? $\Longrightarrow$

Response and predictor, or explanatory, variables.

Having mentioned prediction and inference tasks, we have to ask:

- What are we predicting? $\Longrightarrow$ response variable
- With help of what are we predicting the response variable? $\Longrightarrow$ predictor, or explanatory, variables
- In inference, we are inferring the relationships between...? $\Longrightarrow$ Response and predictor variables.

QuaNTitative or QuaLitative?

Is the variable a) QuaNTitative or b) quaLitative variable? (Numerical or factor? Continuous or categorical?)

5. Person's height → ? quantitative
6. Eye color → ? categorical
7. Test score → ? quantitiative
8. County → ? categorical

QuaNTitative or QuaLitative?

Is the variable a) QuaNTitative or b) quaLitative variable? (Numerical or factor? Continuous or categorical?)

5. Person's height $\rightarrow$?
6. Eye color $\rightarrow$?
7. Test score $\rightarrow$?
8. County $\rightarrow$?

QuaNTitative - something that can be measured and that we can directly perform mathematical operations on (e.g. $3 + 0.2 * Height$)

QuaNTitative or QuaLitative?

Is the variable a) QuaNTitative or b) quaLitative variable? (Numerical or factor? Continuous or categorical?)

5. Person's height $\rightarrow$?
6. Eye color $\rightarrow$?
7. Test score $\rightarrow$?
8. County $\rightarrow$?

QuaNTitative - something that can be measured and that we can directly perform mathematical operations on (e.g. $3 + 0.2 * Height$)

QuaLitative - something that takes on a value of a category or class. Can't perform mathematical operations on them directly (e.g. for color $Color \in \{Red, Green, Blue\}$, what's $3 + 0.2 * Color$?)

QuaNTitative or QuaLitative?

Is the variable a) QuaNTitative or b) quaLitative variable? (Numerical or factor? Continuous or categorical?)

5. Person's height $\rightarrow$?
6. Eye color $\rightarrow$?
7. Test score $\rightarrow$?
8. County $\rightarrow$?

QuaNTitative - something that can be measured and that we can directly perform mathematical operations on (e.g. $3 + 0.2 * Height$)

QuaLitative - something that takes on a value of a category or class. Can't perform mathematical operations on them directly (e.g. for color $Color \in \{Red, Green, Blue\}$, what's $3 + 0.2 * Color$?)

Question: how do we incorporate qualitative variables to perform math operations on them? $\Longrightarrow$

QuaNTitative or QuaLitative?

Is the variable a) QuaNTitative or b) quaLitative variable? (Numerical or factor? Continuous or categorical?)

5. Person's height $\rightarrow$?
6. Eye color $\rightarrow$?
7. Test score $\rightarrow$?
8. County $\rightarrow$?

QuaNTitative - something that can be measured and that we can directly perform mathematical operations on (e.g. $3 + 0.2 * Height$)

QuaLitative - something that takes on a value of a category or class. Can't perform mathematical operations on them directly (e.g. for color $Color \in \{Red, Green, Blue\}$, what's $3 + 0.2 * Color$?)

Question: how do we incorporate qualitative variables to perform math operations on them? $\Longrightarrow$ **Dummy Variables**.

Regression and Classification.

As far as the response variable is concerned, we have:

- Regression task $\rightarrow$ response variable is...

Regression and Classification.

As far as the response variable is concerned, we have:

- Regression task $\rightarrow$ response variable is... qua**NT**itative (continuous, numeric)   Linear
- Classification task $\rightarrow$ response variable is...

Regression and Classification.

As far as the response variable is concerned, we have:

- Regression task → response variable is... qua**NT**itative (continuous, numeric)
- Classification task → response variable is... qua**L**itative (categorical, factor)

In this course, which of the covered methods corresponds to:

9. Regression?

a) Linear Regression

b) Logistic Regression

classification

Logistic ⇒ response is binary

classification models

LDA more than 2 categories

QDA

Regression and Classification.

As far as the response variable is concerned, we have:

- Regression task → response variable is... qua**NT**itative (continuous, numeric)
- Classification task → response variable is... qua**L**itative (categorical, factor)

In this course, which of the covered methods corresponds to:

9. Regression?

   a) Linear Regression   b) Logistic Regression

10. Classification?

   a) Linear Regression   b) Logistic Regression

Simple Linear Regression Example

You're given data on movies' total gross, opening gross, the # of weeks and # of theaters where movie was shown.

Task #1. Assume you are asked to use movies' opening gross to predict their total gross.

- Is it classification or regression? What model do we use?

Simple Linear Regression Example

You're given data on movies' total gross, opening gross, the # of weeks and # of theaters where movie was shown.

Task #1. Assume you are asked to use movies' opening gross to predict their total gross.

- Is it classification or regression? What model do we use?
  Regression, Linear
- What is the model formula for our problem?

Simple Linear Regression Example

You're given data on movies' total gross, opening gross, the # of weeks and # of theaters where movie was shown.

Task #1. Assume you are asked to use movies' opening gross to predict their total gross.

- Is it classification or regression? What model do we use?
- What is the model formula for our problem?

$$\textit{Gross} = \beta_0 + \beta_1 \textit{Opening} + \epsilon, \ \epsilon \sim N(0, \sigma^2)$$

Movies data: Task #1.

```
> summary(lm(Gross~Opening))

Call:
lm(formula = Gross ~ Opening)

Residuals:
    Min      1Q  Median      3Q     Max
-69.567  -4.829  -1.265   4.019 113.830

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.44585    2.31326   0.193    0.848
Opening      2.89977    0.04918  58.957   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.33 on 98 degrees of freedom
Multiple R-squared:  0.9726, Adjusted R-squared:
0.9723
F-statistic:  3476 on 1 and 98 DF,  p-value: < 2.2e-16
```

Estimated model

$\widehat{Gross} = 2.89977 * Opening + 0.44585$

$y = \beta_0 + \beta_1 x + \varepsilon$

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

Movies data: Task #1.

```
> predict(movie.lm,newdata=data.frame(Opening=100),interval="c",level = 0.95)
fit       lwr       upr
1 290.423 281.8443 299.0016
```

(281.8443, 299.0016) $281.8443 \leq \mu_y \leq 299.0$

```
> predict(movie.lm,newdata=data.frame(Opening=100),interval="p",level = 0.95)
fit       lwr       upr
1 290.423 249.1752 331.6707
```

(249.1752, 331.6707)

$249.1752 \leq y \leq 331.6707$

What does a $(281m\$, 299m\$)$ 95% **confidence** interval tell us here?

Movies data: Task #1.

```
> predict(movie.lm,newdata=data.frame(Opening=100),interval="c",level = 0.95)
fit      lwr      upr
1 290.423 281.8443 299.0016
```

(281.8443, 299.0016) $281.8443 \le \mu_y \le 299.0$

```
> predict(movie.lm,newdata=data.frame(Opening=100),interval="p",level = 0.95)
fit      lwr      upr
1 290.423 249.1752 331.6707
```

(249.1752, 331.6707)

$249.1752 \le y \le 331.6707$

What does a $(281m\$, 299m\$)$ 95% **confidence** interval tell us here?

We predict the **average** (μ_y) gross of all movies with an opening gross of $100k\$$ to end up in $(281k\$, 299k\$)$ with 95% confidence.

Movies data: Task #1.

```
> predict(movie.lm,newdata=data.frame(Opening=100),interval="c",level = 0.95)
fit       lwr      upr
1 290.423 281.8443 299.0016
```

(281.8443, 299.0016) $281.8443 \le \mu_y \le 299.0016$

```
> predict(movie.lm,newdata=data.frame(Opening=100),interval="p",level = 0.95)
fit       lwr      upr
1 290.423 249.1752 331.6707
```

(249.1752, 331.6707)

$249.1752 \le y \le 331.6707$

What does a $(281m\$, 299m\$)$ 95% **confidence** interval tell us here?

We predict the **average** (μ_y) gross of all movies with an opening gross of $100k\$$ to end up in $(281k\$, 299k\$)$ with 95% confidence.

What does a $(249m\$, 331m\$)$ 95% **prediction** interval tell us here?

We predict the gross of **any single movie** (y) with an opening gross of $100m\$$ to end up in $(249m\$, 331m\$)$ with 95% confidence.

Movies data: Task #2.

Task # 2 (still *Movies* data): Assume you are asked to use movies' opening gross, # of weeks and # of theaters, to predict their total gross.

- Is it classification or regression? What model do we use?

  Regression, Multivariate liner

Movies data: Task #2.

Task # 2 (still *Movies* data): Assume you are asked to use movies' opening gross, # of weeks and # of theaters, to predict their total gross.

- Is it classification or regression? What model do we use?
- What is the model formula for our problem?

Movies data: Task #2.

Task # 2 (still *Movies* data): Assume you are asked to use movies' opening gross, # of weeks and # of theaters, to predict their total gross.

- Is it classification or regression? What model do we use?
- What is the model formula for our problem?

$$Gross = \beta_0 + \beta_1 Theaters + \beta_2 Opening + \beta_3 Weeks + \epsilon, \; \epsilon \sim N(0, \sigma^2)$$

Movies data: Task #2.

```
movieall.lm=lm(Gross~Theaters+Opening+Weeks)
summary(movieall.lm)

Call:
lm(formula = Gross ~ Theaters + Opening + Weeks)

Residuals:
Min       1Q   Median      3Q      Max
-73.513  -7.733   0.363   4.634  95.983

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.101133   5.403947  -1.314 0.191956
Theaters    -0.002171   0.001850  -1.173 0.243576
Opening      2.904524   0.057292  50.697  < 2e-16 ***
Weeks        1.331971   0.364575   3.653 0.000422 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.15 on 96 degrees of freedom
Multiple R-squared:  0.9762,Adjusted R-squared:  0.9754
F-statistic:  1310 on 3 and 96 DF,  p-value: < 2.2e-16
```

$H_0: \beta_j = 0$, given $\beta_i \neq 0$

$\hat\beta_0$, $\hat\beta_1$, $\hat\beta_2$, $\hat\beta_3$

α $n-3-1$

α

$$\text{Gross} = -7.10113 - 0.002171 \times \text{Theaters} + 2.904524 \times \text{Opening} + 1.331971 \times \text{Weeks}$$

Movies data: Task #2.

We interpret β_j as the average effect of Y (the predictor) of a one unit increase in X_j, **holding all othe prectictors fixed**.

- $\hat{\beta}_2 = 2.905$ This means that for one added million dollars that the movie makes during opening weekend the total gross is predicted to increase on average by \$2.9 million dollars. For a fixed value of the number of theaters and the number of weeks.

- $\hat{\beta}_3 = 1.332$, So for one additional week, the total gross will increase by \$1.33 million dollars for a fixed value of the number of theaters and the opening gross.

Movies data: Task #2.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.101133   5.403947  -1.314 0.191956
Theaters    -0.002171   0.001850  -1.173 0.243576
Opening      2.904524   0.057292  50.697  < 2e-16 ***
Weeks        1.331971   0.364575   3.653 0.000422 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0: \beta_0 = 0$
$H_0: \beta_1 = 0$
$H_0: \beta_2 = 0$
$H_0: \beta_3 = 0$

- Notice for tesing $H_0 : \beta_1 = 0$, P-value = 0.24357.
- This is testing if we need the variable Theaters if Opening and Weeks are in the model.
- Thus Theaters is not needed to predict the total Gross for movies.

Movies data: Task #2.

- Confidence interval

```
> predict(moviet.lm,newdata = data.frame(Opening=90,Weeks=15),interval="c",level=0.95)
       fit      lwr      upr
1 266.3005 258.5255 274.0755
```

This means we predict the **average** total gross for a movie that has a opeining weekend gross of $90 million dollars and has been in the theaters for 15 weeks to be in [266.31, 258.52] with 95% confidence.

- Prediction interval

```
> predict(moviet.lm,newdata = data.frame(Opening=90,Weeks=15),interval="p",level=0.95)
       fit      lwr      upr
1 266.3005 227.4304 305.1705
```

This means we predict the total gross for a movie that has a opeining weekend gross of $90 million dollars and has been in the theaters for 15 weeks to be in [227.43, 305.17] with 95% confidence.

Movies data: Task #3.

Example: Assume you are asked to describe a relationship between movies' opening gross and the # of theaters**.**

Question: If asked to describe a relationship between two quaNTitative variables, what do we, as extremely promising data scientists, do **first**?

Movies data: Task #3.

Example: Assume you are asked to describe a relationship between movies' opening gross and the # of theaters.

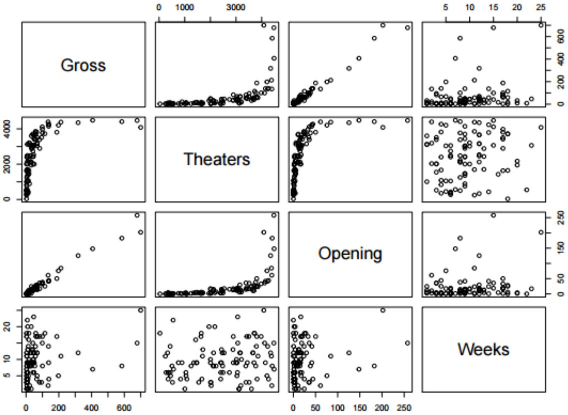
Question: If asked to describe a relationship between two quaNTitative variables, what do we, as extremely promising data scientists, do **first**?

Answer:

P-L-O-T (or V-I-S-U-A-L-I-Z-E).
T-H-E D-A-T-A

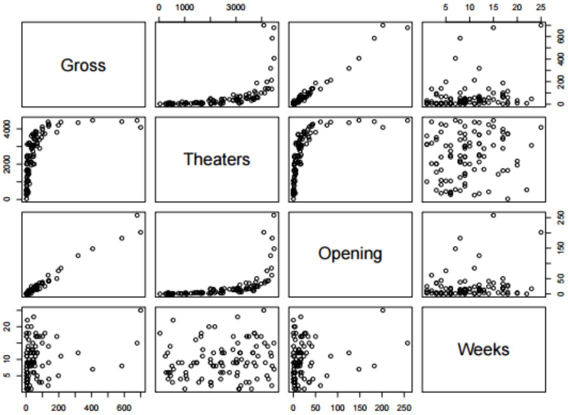
Movies data: Task #3.

```
pairs(movies[,3:6])
```



Movies data: Task #3.

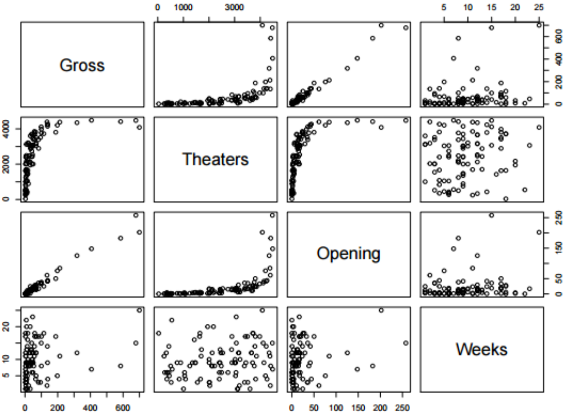
```
pairs(movies[,3:6])
```



Gross and theaters show a clear non-linear pattern. How do we deal with that? $\rightarrow$

Movies data: Task #3.

```
pairs(movies[,3:6])
```



Gross and theaters show a clear non-linear pattern. How do we deal with that? $\rightarrow$ **Polynomial regression**.

Movies data: Task #3.

Model for quadratic polynomial regression of *Gross* on *Theaters* is

$$Gross = \beta_0 + \beta_1 Theaters + \beta_2 Theaters^2, \; \epsilon \sim N(0, \sigma^2)$$

In *R* it can be carried out as:

```
> lm.polynom <- lm(Gross ~ poly(Theaters,2), data=movies)
> summary(lm.polynom)
...
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)             65.51       8.69   7.538 2.54e-11 ***
poly(Theaters, 2)1    675.15      86.90   7.769 8.27e-12 ***
poly(Theaters, 2)2    537.51      86.90   6.186 1.47e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
...
```

Is the quadratic relationship significant?

BreastCancer data.

Example. You are given multiple breast cancer tissue samples that are labeled as either malignant ($Class = 1$) or benign ($Class = 0$). We want to use cell shape ($Cell.shape$) to predict class labels.

- Is it classification or regression? What model do we use?

Classification, Logistic

BreastCancer data.

Example. You are given multiple breast cancer tissue samples that are labeled as either malignant ($Class = 1$) or benign ($Class = 0$). We want to use cell shape ($Cell.shape$) to predict class labels.

- Is it classification or regression? What model do we use?
- What is the model formula for our problem?

BreastCancer data.

Example. You are given multiple breast cancer tissue samples that are labeled as either malignant ($Class = 1$) or benign ($Class = 0$). We want to use cell shape ($Cell.shape$) to predict class labels.

- Is it classification or regression? What model do we use?
- What is the model formula for our problem?

  While in linear regression we can model the response Y directly:

  $$Y = \beta_0 + \beta_1 X + ...$$

  in logistic regression ($Y = 0/1$)...

BreastCancer data.

Example. You are given multiple breast cancer tissue samples that are labeled as either malignant ($Class = 1$) or benign ($Class = 0$). We want to use cell shape ($Cell.shape$) to predict class labels.

- Is it classification or regression? What model do we use?
- What is the model formula for our problem?

  While in linear regression we can model the response Y directly:

  $$Y = \beta_0 + \beta_1 X + ...$$

  in logistic regression ($Y = 0/1$)... we need to do transformations:

BreastCancer data.

Example. You are given multiple breast cancer tissue samples that are labeled as either malignant ($Class = 1$) or benign ($Class = 0$). We want to use cell shape ($Cell.shape$) to predict class labels.

- Is it classification or regression? What model do we use?
- What is the model formula for our problem?

  While in linear regression we can model the response Y directly:

  $$Y = \beta_0 + \beta_1 X + ...$$

  in logistic regression ($Y = 0/1$)... we need to do transformations:

  - First, let $p(X) = P(Y = 1|X)$ - we will model probability of $Y = 1$.

BreastCancer data.

Example. You are given multiple breast cancer tissue samples that are labeled as either malignant ($Class = 1$) or benign ($Class = 0$). We want to use cell shape ($Cell.shape$) to predict class labels.

- Is it classification or regression? What model do we use?
- What is the model formula for our problem?

  While in linear regression we can model the response Y directly:

  $$Y = \beta_0 + \beta_1 X + ...$$

  in logistic regression ($Y = 0/1$)... we need to do transformations:

  - First, let $p(X) = P(Y = 1|X)$ - we will model probability of $Y = 1$.
  - Can we do $p(X) = \beta_0 + \beta_1 X + ..$?

BreastCancer data.

Example. You are given multiple breast cancer tissue samples that are labeled as either malignant ($Class = 1$) or benign ($Class = 0$). We want to use cell shape (*Cell.shape*) to predict class labels.

- Is it classification or regression? What model do we use?
- What is the model formula for our problem?

  While in linear regression we can model the response Y directly:

  $$Y = \beta_0 + \beta_1 X + ...$$

  in logistic regression ($Y = 0/1$)... we need to do transformations:

  - First, let $p(X) = P(Y = 1|X)$ - we will model probability of $Y = 1$.
  - Can we do $p(X) = \beta_0 + \beta_1 X + ..$? No, the left side is stuck $\in [0, 1]$.

BreastCancer data.

Example. You are given multiple breast cancer tissue samples that are labeled as either malignant ($Class = 1$) or benign ($Class = 0$). We want to use cell shape ($Cell.shape$) to predict class labels.

- Is it classification or regression? What model do we use?
- What is the model formula for our problem?

  While in linear regression we can model the response Y directly:

$$Y = \beta_0 + \beta_1 X + ...$$

  in logistic regression ($Y = 0/1$)... we need to do transformations:

  - First, let $p(X) = P(Y = 1|X)$ - we will model probability of $Y = 1$.
  - Can we do $p(X) = \beta_0 + \beta_1 X + ..$? No, the left side is stuck $\in [0, 1]$.
  - $p(X) \in [0, 1] \implies$

BreastCancer data.

Example. You are given multiple breast cancer tissue samples that are labeled as either malignant ($Class = 1$) or benign ($Class = 0$). We want to use cell shape ($Cell.shape$) to predict class labels.

- Is it classification or regression? What model do we use?
- What is the model formula for our problem?

  While in linear regression we can model the response Y directly:

  $$Y = \beta_0 + \beta_1 X + ...$$

  in logistic regression ($Y = 0/1$)... we need to do transformations:

  - First, let $p(X) = P(Y = 1|X)$ - we will model probability of $Y = 1$.
  - Can we do $p(X) = \beta_0 + \beta_1 X + ..$? No, the left side is stuck $\in [0, 1]$.
  - $p(X) \in [0, 1] \implies \frac{p(X)}{1-p(X)} \in (0, \infty) \implies$

BreastCancer data.

Example. You are given multiple breast cancer tissue samples that are labeled as either malignant ($Class = 1$) or benign ($Class = 0$). We want to use cell shape ($Cell.shape$) to predict class labels.

- Is it classification or regression? What model do we use?
- What is the model formula for our problem?

  While in linear regression we can model the response Y directly:

  $$Y = \beta_0 + \beta_1 X + ...$$

  in logistic regression ($Y = 0/1$)... we need to do transformations:

  - First, let $p(X) = P(Y = 1|X)$ - we will model probability of $Y = 1$.
  - Can we do $p(X) = \beta_0 + \beta_1 X + ..$? No, the left side is stuck $\in [0, 1]$.
  - $p(X) \in [0, 1] \implies \frac{p(X)}{1-p(X)} \in (0, \infty) \implies \log\left(\frac{p(X)}{1-p(X)}\right) \in (-\infty, \infty)$

BreastCancer data.

Example. You are given multiple breast cancer tissue samples that are labeled as either malignant ($Class = 1$) or benign ($Class = 0$). We want to use cell shape ($Cell.shape$) to predict class labels.

- Is it classification or regression? What model do we use?
- What is the model formula for our problem?

  While in linear regression we can model the response Y directly:

$$Y = \beta_0 + \beta_1 X + ...$$

  in logistic regression ($Y = 0/1$)... we need to do transformations:

  - First, let $p(X) = P(Y = 1|X)$ - we will model probability of $Y = 1$.
  - Can we do $p(X) = \beta_0 + \beta_1 X + ..$? No, the left side is stuck $\in [0, 1]$.
  - $p(X) \in [0, 1] \implies \frac{p(X)}{1-p(X)} \in (0, \infty) \implies \log\left(\frac{p(X)}{1-p(X)}\right) \in (-\infty, \infty)$
  - formula $\log(\frac{p(X)}{1-p(X)})$ is also known as logit.

BreastCancer data.

Denoting $p(Class = Malignant \mid X) = p(X)$, the model formula is:

$$\star \log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 Cell.shape$$

$$\star \; P(X) = \frac{\exp\{\beta_0 + \beta_1 \, cell.shape\}}{1 + \exp\{\beta_0 + \beta_1 \, cell.shape\}}$$

$$\cancel{p(x) = \beta_0 + \beta_1 \, cell.shape}$$

BreastCancer data.

Denoting $p(Class = Malignant \mid X) = p(X)$, the model formula is:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 Cell.shape$$

```
summary(glm(Class ~ Cell.shape, family="binomial", data = bc))

Call:
glm(formula = Class ~ Cell.shape, family = "binomial", data = bc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6383  -0.2219  -0.2219   0.0517   2.7263

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.1645     0.3865  -13.36   <2e-16 ***
Cell.shape    1.4727     0.1205   12.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0: \beta_1 = 0$

Interpretation: $\hat{\beta}_1 = 1.47 \implies$ as cell shape increases, the probability of tissue being *Malignant* also **increases**.

Titanic data.

Example. Given the survival (Yes/No) and gender (Male/Female) data of Titanic passengers, use gender to explain the survival outcome.

Titanic data.

Example. Given the survival (Yes/No) and gender (Male/Female) data of Titanic passengers, use gender to explain the survival outcome.

- How do we represent gender factor variable in the model?

Titanic data.

Example. Given the survival (Yes/No) and gender (Male/Female) data of Titanic passengers, use gender to explain the survival outcome.

- How do we represent gender factor variable in the model?

  Dummy variable(s). E.g. here: $x_{Sex} = \begin{cases} 0, Sex = Female, \\ 1, Sex = Male \end{cases}$

Titanic data.

Example. Given the survival (Yes/No) and gender (Male/Female) data of Titanic passengers, use gender to explain the survival outcome.

- How do we represent gender factor variable in the model?

  Dummy variable(s). E.g. here: $x_{Sex} = \begin{cases} 0, Sex = Female, \\ 1, Sex = Male \end{cases}$

- What is the model formula?

Titanic data.

Example. Given the survival (Yes/No) and gender (Male/Female) data of Titanic passengers, use gender to explain the survival outcome.

- How do we represent gender factor variable in the model?

  Dummy variable(s). E.g. here: $x_{Sex} = \begin{cases} 0, Sex = Female, \\ 1, Sex = Male \end{cases}$

- What is the model formula? Letting $p(Surv = Yes \mid X) = p(X)$:

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 x_{Sex}$$

$$P(X) = \begin{cases} \dfrac{\exp\{\beta_0\}}{1 + \exp\{\beta_0\}} & \text{if } Sex = female \\ \dfrac{\exp\{\beta_0 + \beta_1\}}{1 + \exp\{\beta_0 + \beta_1\}} & \text{if } Sex = male \end{cases}$$

Titanic data.

```
> summary(glm(ifelse(Survived=="Yes",1,0)~Sex,
                      family="binomial",
                       data=Titanic_tab,weights=Count))
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.8127     0.1070   7.592 3.14e-14 ***
SexMale      -2.1255     0.1221 -17.404  < 2e-16 ***
...
```

Interpretation: there is a significant gender effect on the survival outcome. Males had lower survival probability (or "the logit of survival probability for males was on average 2.1255 lower than for females").

NOTE: For FACTOR VARIABLES, **DON'T** SAY "Per 1 unit increase in [factor variable], the [response or logit probability] on average decreases by 2.12..". It applies to both **linear** and **logistic** regression.

Predicting Well?

Confusion Matrix

	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	Correct true negatives	Incorrect false positives
$Y = 1$	Incorrect false negative	Correct true positives

$$\text{error rate} = \frac{(\hat{Y}=1 \cap Y=0) + (\hat{Y}=0 \cap Y=1)}{\text{Total}}$$

$$\text{sensitivity} = \frac{(\hat{Y}=1 \cap Y=1)}{Y=1}$$

$$\text{Specifity} = \frac{\hat{Y}=0 \cap Y=0}{Y=0}$$

$$\text{success rate} = 1 - \text{error rate}$$

Linear Discriminate Analysis

- Classification problem
- Can be used for qualitative response (categorical) with more than two categories.
- Uses a prior probability, assume a Normal distribution.
- Posterior probability $P(Y = k|X = x)$, then the category with the highest posterior posterior is the predicted category.

```
> library(MASS)
    > cars.lda = lda(cylinders ~ mtcars$mpg + mtcars$hp)
> cars.lda
Call:
lda(cylinders ~ mtcars$mpg + mtcars$hp)

Prior probabilities of groups:
4       6       8
0.34375 0.21875 0.43750

Group means:
mtcars$mpg mtcars$hp
4   26.66364  82.63636
6   19.74286 122.28571
8   15.10000 209.21429

Coefficients of linear discriminants:
LD1          LD2
mtcars$mpg -0.2020452 0.25260148
mtcars$hp   0.0157379 0.02254518

Proportion of trace:
LD1    LD2
0.9694 0.0306
> cars.pred = predict(cars.lda)
> table(cylinders,cars.pred$class)

cylinders  4  6  8
4          9  2  0
6          0  6  1
8          0  0 14
```

error rate $= \frac{3}{32} = 0.09$

$9+2+4+1+14=32$