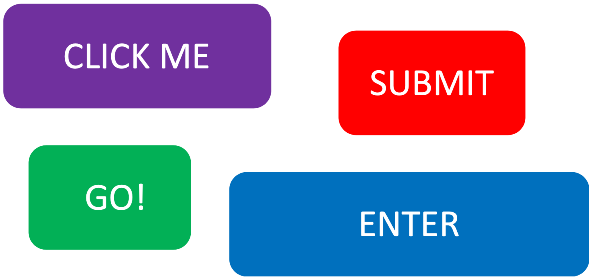
Experiments with Multiple Design Factors

- We now consider the design and analysis of an experiments consisting of multiple conditions arising from multiple design factors.
    - This is often colloquially referred to as "multivariate testing" (MVT)

- Canonical MVT button test:



- Some additional, more tangible, examples:
    - Etsy
    - Netflix
    - Airbnb

- This week we describe how to design and analyze experiments that efficiently investigate multiple design factors

- **Goals:**

1. Determine which condition is optimal
    - But now a condition is defined by a specific combination of the levels of *multiple* design factors

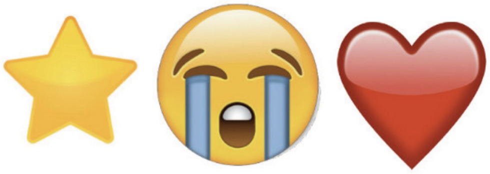
2. Determine *which* factors are influential and understand *how* the factors influence the response

The Factorial Approach

- The key to multi-factor experiments is to *efficiently* investigate different combinations of the factor levels

- **One-factor-at-a-time approach:** (greedy)
    - Sequence of experiments, each with just one factor being varied
    - The winning level of this factor is retained
    - Follow-up experiments manipulate some other factor, while the previous ones are held fixed at their optimal levels

Example 1: Twitter Hearts vs Stars



- In 2015 Twitter changed "favoriting" a tweet (expressed as stars) to "liking" a tweet (expressed as hearts) and the internet was pissed
- In line with Twitter's "test everything" motto, this decision came about as a result of experimentation
- A hypothetical experiment that could have lead to this decision might involve two factors at two levels

  DF1 (shape): {star, heart}

  DF2 (color): {Yellow, Red}

- A one-factor-at-a-time approach might look like this:

  Yellow star < Yellow heart < Red heart

  ↑ exp 1 ↑ exp 2

  missed out the Red Star.

- **Example 2: Etsy Search Bar**
  - Check it out

- **The Factorial approach:**
  - Experimental conditions are defined as every unique combination of the design factors' levels
  - In the **Twitter** example, the factorial experiment would have looked like this:
  - In the **Etsy** example, the factorial experiment would have looked like this:

  - **Advantage:** it explores every possible condition

Better understanding of relationship btwn. response & factors compared to OF. A.T. experiments.

  - **Disadvantage:** it explores every possible condition

large space to search

  - As long as we choose our factors and their levels thoughtfully, the advantages outweigh the disadvantages

- **Main effects:** The main effect of factor A, represents the change in the response variable produced by a change in that factor

- **Interaction effects:** If the main effect of factor A depends on the level of some other factor B, we say that factors A and B interact

- From a practical perspective it is critical to quantify both types of effects

Designing Factorial Experiments

- Conceptually, the design of a factorial experiment is simple
  - Pick your metric of interest and define the corresponding response variable
  - Pick your design factors
  - Pick their levels
  - Define your experimental conditions
  - Determine your sample sizes

- **Button Example:** Suppose you have $K = 3$ factors, *colour* (red, blue), *phrase* (“Continue”, “Go”), and *size* (small, medium, large). These factors therefore have $m_1 = 2$, $m_2 = 2$, and $m_3 = 3$ levels respectively

- In general, a factorial experiment with K factors requires $m = \underbrace{m_1 m_2 \cdots m_K}$ conditions, where m_k is the number of levels of design factor k.

Smallest factorial design is 2^K levels (each factor 2 levels)

- As the number of factors and levels increase, the size of the experiment can get unmanageably large
  - As such, we want to pick our factors and levels thoughtfully

Keep it simple }

1. Don't investigate factors that are highly correlated
2. Try not choose similar levels
3. Don't choose factors that are hard to manipulate

- Once the factors, levels, and hence experimental conditions have been established, experimental units must be randomized to each of the m conditions
  - The number of experimental units assigned to each condition n_j, $j = 1, 2, \ldots, m$, can be determined by sample size calculations associated with two-sample tests

and make sure to account for the multiple comparison problem

Analyzing Factorial Experiments

- In order to determine which condition is optimal we use pairwise tests

- In order to determine which factors are influential, and to quantify this influence we use regression

- Whether it's a linear or logistic regression, we use a linear predictor which contains the following terms:
  - An intercept
  - Main effect terms
  - Two-factor interaction terms
  - Three-factor interaction terms

    ⋮

  - K-factor interaction terms

- **Button Example:** With $K = 3$ factors with $m_1 = 2$, $m_2 = 2$ and $m_3 = 3$ levels the required linear predictor will contain:
  - Main effect terms, and
  - Two-factor interaction terms, and
  - Three-factor interaction terms
  - The linear predictor is given by:
  - Hypotheses concerning the main effects will therefore involve $\beta_1, \beta_2, \beta_3, \beta_4$
  - Hypotheses concerning the two-factor interactions will involve $\beta_5, \beta_6, \beta_7, \beta_8, \beta_9$
  - Hypotheses concerning the three-factor interactions will involve β_{10}, β_{11}

- In general, hypotheses of these sort will be performed by comparing full versus reduced models via partial F-tests (in the case of linear regression) and likelihood ratio tests (in the case of logistic regression)

- **Another Example:** Suppose Factor A has $m_1 = 5$ levels, Factor B has $m_2 = 2$ levels and Factor C has $m_3 = 3$ levels.
  - **Factor A** will be represented in a regression model by $m_1 - 1 = 4$ indicator variables:

    and $m_1 - 1 = 4$ corresponding β's:
  - Thus the *main effect* of Factor A is composed of 4 *terms* in the model.
  - To determine the significance of the main effect of Factor A, we test

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

  - **Factor B** will be represented in a regression model by $m_2 - 1 = 1$ indicator variable:

    and $m_2 - 1 = 1$ corresponding β:
  - Thus the *main effect* of Factor B is composed of 1 *term* in the model.
  - To determine the significance of the main effect of Factor B, we test

$$H_0 : \beta_5 = 0$$

  - **Factor C** will be represented in a regression model by $m_3 - 1 = 2$ indicator variables:

    and $m_3 - 1 = 2$ corresponding β's:
  - Thus the *main effect* of Factor C is composed of 2 *terms* in the model.
  - To determine the significance of the main effect of Factor C, we test

$$H_0 : \beta_6 = \beta_7 = 0$$

  - **Note:** These three hypotheses are relevant only in the context of the *main effect model* which has linear predictor

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$

- The interaction effect between *two* factors is represented by two-way products of the indicator variables corresponding to the main effects of the two factors.
  - The **A:B interaction** *effect* is composed of the $(m_1 - 1) \times (m_2 - 1) = 4 \times 1 = 4$ *terms* resulting from the two-way products between Factor A's and Factor B's indicator variables:

    with corresponding β's:

  - The significance of the A:B interaction effect is determined by testing

$$H_0 : \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$$

  - The **A:C interaction** *effect* is composed of the $(m_1 - 1) \times (m_3 - 1) = 4 \times 2 = 8$ *terms* resulting from the two-way products between Factor A's and Factor C's indicator variables:

    with corresponding β's:

  - The significance of the A:C interaction effect is determined by testing

$$H_0 : \beta_{12} = \beta_{13} = \beta_{14} = \beta_{14} = \beta_{15} = \beta_{16} = \beta_{17} = \beta_{18} = \beta_{19} = 0$$

  - The **B:C interaction** *effect* is composed of the $(m_2 - 1) \times (m_3 - 1) = 1 \times 2 = 2$ *terms* resulting from the two-way products between Factor B's and Factor C's indicator variables:

    with corresponding β's:

  - The significance of the B:C interaction effect is determined by testing

$$H_0 : \beta_{20} = \beta_{21} = 0$$

- The interaction effect between *three* factors is represented by three-way products of the indicator variables corresponding to the main effects of the three factors.
  - The **A:B:C interaction** *effect* is composed of the $(m_1 - 1) \times (m_2 - 1) \times (m_3 - 1) = 4 \times 1 \times 2 = 8$ *terms* resulting from the three-way products between Factor A's, Factor B's, and Factor C's indicator variables:

    with corresponding β's:

  - The significance of the A:B:C interaction effect is determined by testing

$$H_0 : \beta_{22} = \beta_{23} = \beta_{24} = \beta_{25} = \beta_{26} = \beta_{27} = \beta_{28} = \beta_{29} = 0$$

  - **Note:** The hypotheses concerning the significance of interaction effects are relevant only in the context of the *full model* which has linear predictor

$$\beta_0 +$$

$$\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 +$$

$$\beta_8 x_1 x_5 + \beta_9 x_2 x_5 + \beta_{10} x_3 x_5 + \beta_{11} x_4 x_5 +$$

$$\beta_{12} x_1 x_6 + \beta_{13} x_2 x_6 + \beta_{14} x_3 x_6 + \beta_{15} x_4 x_6 + \beta_{16} x_1 x_7 + \beta_{17} x_2 x_7 + \beta_{18} x_3 x_7 + \beta_{19} x_4 x_7 +$$

$$\beta_{20} x_5 x_6 + \beta_{21} x_5 x_7 +$$

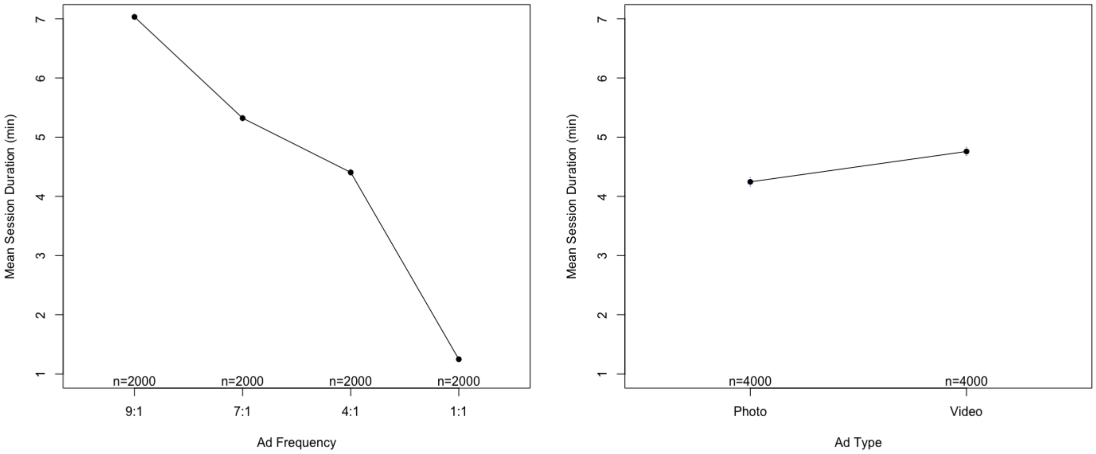
$$\beta_{22} x_1 x_5 x_6 + \beta_{23} x_1 x_5 x_7 + \beta_{24} x_2 x_5 x_6 + \beta_{25} x_2 x_5 x_7 + \beta_{26} x_3 x_5 x_6 + \beta_{27} x_3 x_5 x_7 + \beta_{28} x_4 x_5 x_6 + \beta_{29} x_4 x_5 x_7$$

- **ALL** of the tests discussed here are carried by either a **partial F-test** or a **likelihood ratio test**.

Continuous Response

- We illustrate the topics discussed in this section in the context of an **Instagram Ad** example.
- Suppose that you are a data scientist at Instagram, and you are interested in running an experiment to learn about how user engagement is influenced by ad frequency and ad type.
- Suppose that ad frequency has levels {*9:1*, *7:1*, *4:1*, *1:1*} corresponding to ad frequencies of 1 in 10, 1 in 8, 1 in 5, and every other.
- Suppose also that ad type is a second design factor with levels {*photo*, *video*}.
- We will consider here the factorial experiment that considers every combination of these two factors' levels.

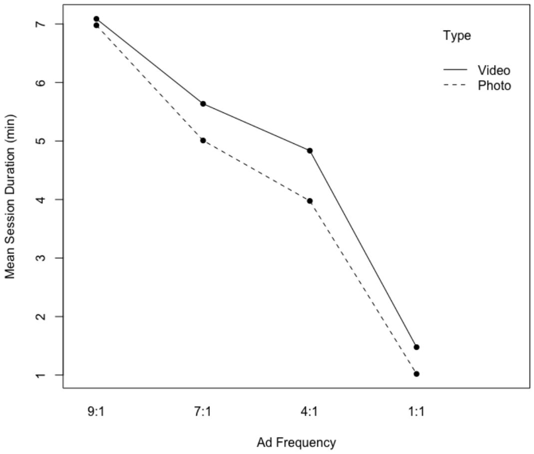
- Assume $n = 1000$ users are randomly assigned to each of these $m = 8$ conditions, and on each user we measure the length of time they engage with the app (in minutes).

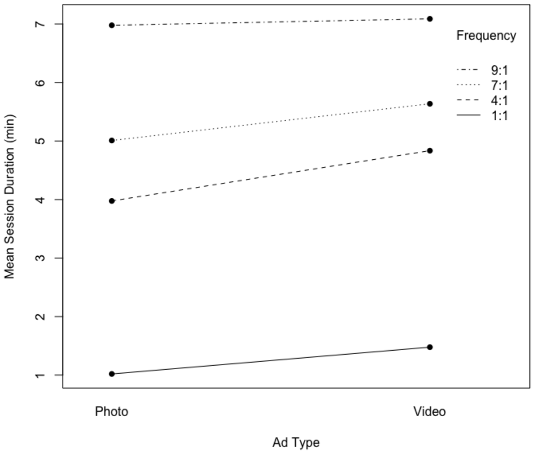


- The resulting data was used to create the following **main effect plots:**

- **Important:**
  - Discussing main effects can be uninformative and potentially misleading if there is a signficant interaction between the factors
  - In the presence of a significant interaction effect, it no longer makes sense to discuss the main effect of a factor in isolation, because doing so ignores the fact that this effect changes depending on the level of another factor

- We can evaluate the presence of such interaction by studying **interaction effect plots:**
- Non-parallel line segments on these plots would indicate the presence of an interaction since this would correspond to the main effect of one factor depending on the levels of the other factor.





- To formally evaluate whether these main effects and interaction effects are significant we fit the following linear regression model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i1} x_{i4} + \beta_6 x_{i2} x_{i4} + \beta_7 x_{i3} x_{i4} + \epsilon_i$$

  where the x's are indicator variables

- The expected response in each condition, according to this model is:

		Ad Type	
		Photo	Video
Freq.	9:1	$E[Y_i \vert x_{i1} = x_{i2} = x_{i3} = 0, x_{i4} = 0] = \beta_0$	$E[Y_i \vert x_{i1} = x_{i2} = x_{i3} = 0, x_{i4} = 1] = \beta_0 + \beta_4$
	7:1	$E[Y_i \vert x_{i1} = 1, x_{i4} = 0] = \beta_0 + \beta_1$	$E[Y_i \vert x_{i1} = 1, x_{i4} = 1] = \beta_0 + \beta_1 + \beta_4 + \beta_5$
	4:1	$E[Y_i \vert x_{i2} = 1, x_{i4} = 0] = \beta_0 + \beta_2$	$E[Y_i \vert x_{i2} = 1, x_{i4} = 1] = \beta_0 + \beta_2 + \beta_4 + \beta_6$
	1:1	$E[Y_i \vert x_{i3} = 1, x_{i4} = 0] = \beta_0 + \beta_3$	$E[Y_i \vert x_{i3} = 1, x_{i4} = 1] = \beta_0 + \beta_3 + \beta_4 + \beta_7$

β_4 is not an interaction coefficient

- Clearly, a formal test of

$$H_0\colon\ \beta_5 = \beta_6 = \beta_7 = 0 \text{ vs. } H_A\colon\ \beta_j \neq 0$$

  for $j = 5, 6, 7$ would evaluate the significance of the interaction effect

  - If we reject H_0, any conclusions regarding the effect of one factor must be made in the context of the levels of the other factor

  - If we do not reject H_0, the interaction terms can be removed from the model yielding the following simplified **main effects** model:

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

    which can be used to evaluate the significance of the main effect of each factor

- The expected response in each condition, according to the main effects model is: (assume no interaction)

		Ad Type	
		Photo	Video
Freq.	9:1	$E[Y_i \mid x_{i1} = x_{i2} = x_{i3} = 0, x_{i4} = 0] = \beta_0$	$E[Y_i \mid x_{i1} = x_{i2} = x_{i3} = 0, x_{i4} = 1] = \beta_0 + \beta_4$
	7:1	$E[Y_i \mid x_{i1} = 1, x_{i4} = 0] = \beta_0 + \beta_1$	$E[Y_i \mid x_{i1} = 1, x_{i4} = 1] = \beta_0 + \beta_1 + \beta_4$
	4:1	$E[Y_i \mid x_{i2} = 1, x_{i4} = 0] = \beta_0 + \beta_2$	$E[Y_i \mid x_{i2} = 1, x_{i4} = 1] = \beta_0 + \beta_2 + \beta_4$
	1:1	$E[Y_i \mid x_{i3} = 1, x_{i4} = 0] = \beta_0 + \beta_3$	$E[Y_i \mid x_{i3} = 1, x_{i4} = 1] = \beta_0 + \beta_3 + \beta_4$

- The hypothesis

  $$H_0\colon\ \beta_1 = \beta_2 = \beta_3 = 0 \text{ vs. } H_A\colon\ \beta_j \neq 0$$

  for $j = 1, 2, 3$ tests whether ad frequency is a significant factor

- The hypothesis

  $$H_0\colon\ \beta_4 = 0 \text{ vs. } H_A\colon\ \beta_4 \neq 0$$

  tests whether ad type is a significant factor

- **But remember:** these tests and the interpretation of main effects are only appropriate in the absence of interaction.

- Each of these null hypotheses generates a **reduced model** with fewer terms relative to a **full model** with all terms – we compare them using using partial F-tests associated with an analysis of variance.

- Output from the relevant partial F-tests is shown below

```
Analysis of Variance Table

Model 1: Time ~ Frequency + Type
Model 2: Time ~ Frequency * Type
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   7995 6522.2
2   7992 6372.9  3    149.27 62.398 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

concl.:
there are interactions

```
Analysis of Variance Table

Model 1: Time ~ Frequency
Model 2: Time ~ Frequency + Type
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   7996 7049.5
2   7995 6522.2  1    527.34 646.43 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

concl:
Type is important

```
Analysis of Variance Table

Model 1: Time ~ Type
Model 2: Time ~ Frequency + Type
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)
1   7998 41875
2   7995  6522  3     35353 14445 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

concl.:
freq. is important

- Conclusions:

Factorial Experiments – Binary Response

- The informal and formal evaluation of main and interaction effects can be performed in the context of a binary response variable as well.
  - Main effect and interaction effect plots are based on observed proportions
  - Logistic regression is used instead of ordinary linear regression

- The structure of the linear predictor is identical to what we have discussed in general

- For instance, if the Instagram experiment from the previous class had a binary response instead, the relevant logistic regression model would be

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i1} x_{i4} + \beta_6 x_{i2} x_{i4} + \beta_7 x_{i3} x_{i4}$$

  where the x's are the indicator variables as defined previously.

- Interest lies in determining whether subsets of the β's are equal to zero to evaluate the significance of various main and interaction effects
  - We use **likelihood ratio tests** for the comparison of full and reduced logistic regression models
  - The test statistic for the LRT is:

$$\begin{aligned} t &= 2 \times \log\left(\frac{\text{Likelihood}_{\text{Full Model}}}{\text{Likelihood}_{\text{Reduced Model}}}\right) \\ &= 2 \times [\text{Log-Likelihood}_{\text{Full Model}} - \text{Log-Likelihood}_{\text{Reduced Model}}] \end{aligned}$$

  - The p-value for this test is calculated as

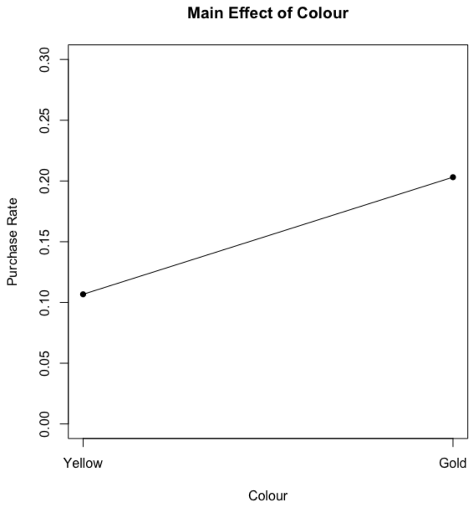
$$\text{p-value} = P(T \geq t)$$

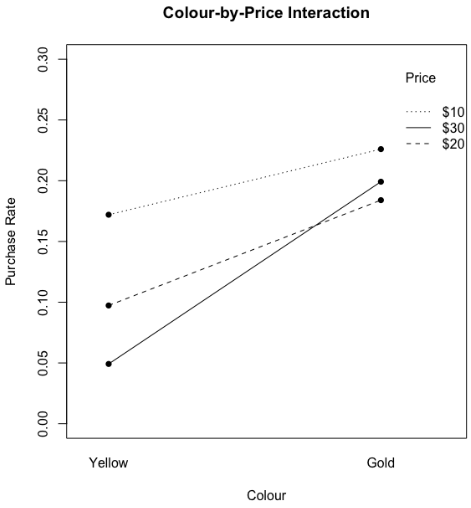
    where $T \sim \chi^2_{\textcircled{(\nu)}}$

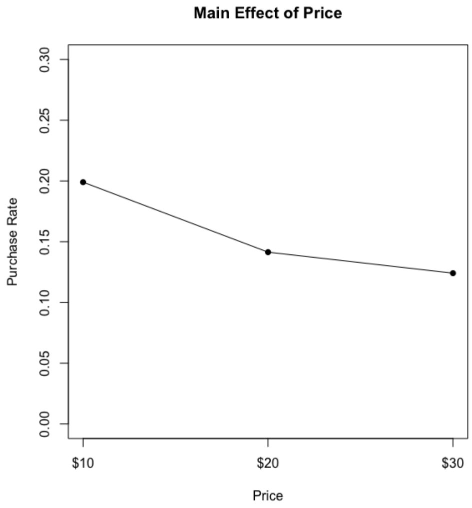
The TinyCo Example

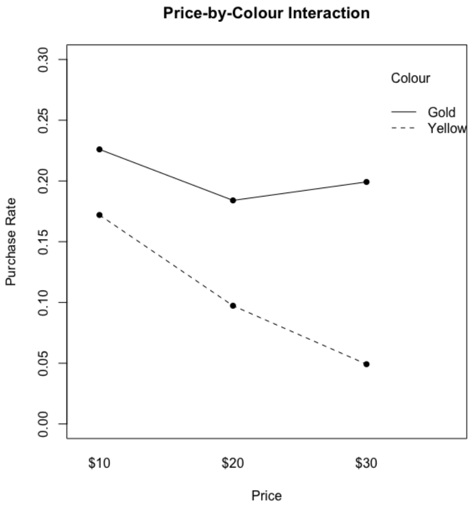
TinyCo is a mobile video game studio that develops the Tiny Zoo game. In this game users own zoos and collect animals to put in their zoos. An experiment is performed in which a new animal, the “bananimal”, is released for purchase as a part of the Super Sweet Series. Interest lies in understanding the relationship between conversion (purchase rate) and two factors: the bananimal’s colour (yellow or gold) and the bananimal’s price ($10, $20, or $30 of in-game currency). A factorial experiment with 6 conditions was performed to investigate these relationships. A summary of the data resulting from this experiment is shown below.

Condition	Sample Size	Purchase Rate
$10 + Yellow	500	0.1720
$20 + Yellow	483	0.0973
$30 + Yellow	488	0.0492
$10 + Gold	500	0.2260
$20 + Gold	500	0.1840
$30 + Gold	487	0.1992


Main Effect of Colour


Colour-by-Price Interaction


Main Effect of Price


Price-by-Colour Interaction

- What do the main effect plots tell us?

- What do the interaction effect plots tell us?

- To formally analyze this data we fit the *full* logistic regression model with linear predictor

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3}$$

- We test the significance of the interaction effect via the hypothesis

$$H_0 : \beta_4 = \beta_5 = 0 \text{ vs. } H_A : \exists\ j \in \{4, 5\} \text{ s.t. } \beta_j \neq 0$$

- This involves a comparison between the full model and the reduced *main effects* model with linear predictor

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

- We can also test the main effect of colour with $H_0 : \beta_1 = 0$ in the context of the main effects model:

- We can also test the main effect of price with $H_0 : \beta_2 = \beta_3 = 0$ in the context of the main effects model:

- So what have we learned about the influence of these factors?

- And which condition was optimal?

Optional Exercises:

- R Analysis: 10, 20, 25
- Communication: 1(e)