Contents

review

- big picture of applied stats: see 36200 image idk
- we have statistics ($\overline{x}, \hat{p}, \ldots$) and standard error ($\mathrm{SE}_{\overline{x}}, \mathrm{SE}_{\hat{p}}, \ldots$)
- population: literally everyone, hard to measure
- sample: subset of population
- parameter: perfect summary (e.g. mean height)
- statistic: measurable summary (e.g. mean height of sample)
- stderr of stat: typical variation due to random sampling.
  - diff error formulae for each stat.
  - this course: simply calc with software
- inference: give estimate and measure of how far off it might be
  - if statistic is random and sampling distribution known, we have probabilistic inference; can get p-value or margin or err

1 variable EDA

- categorical
  - bar graph
  - percent summaries

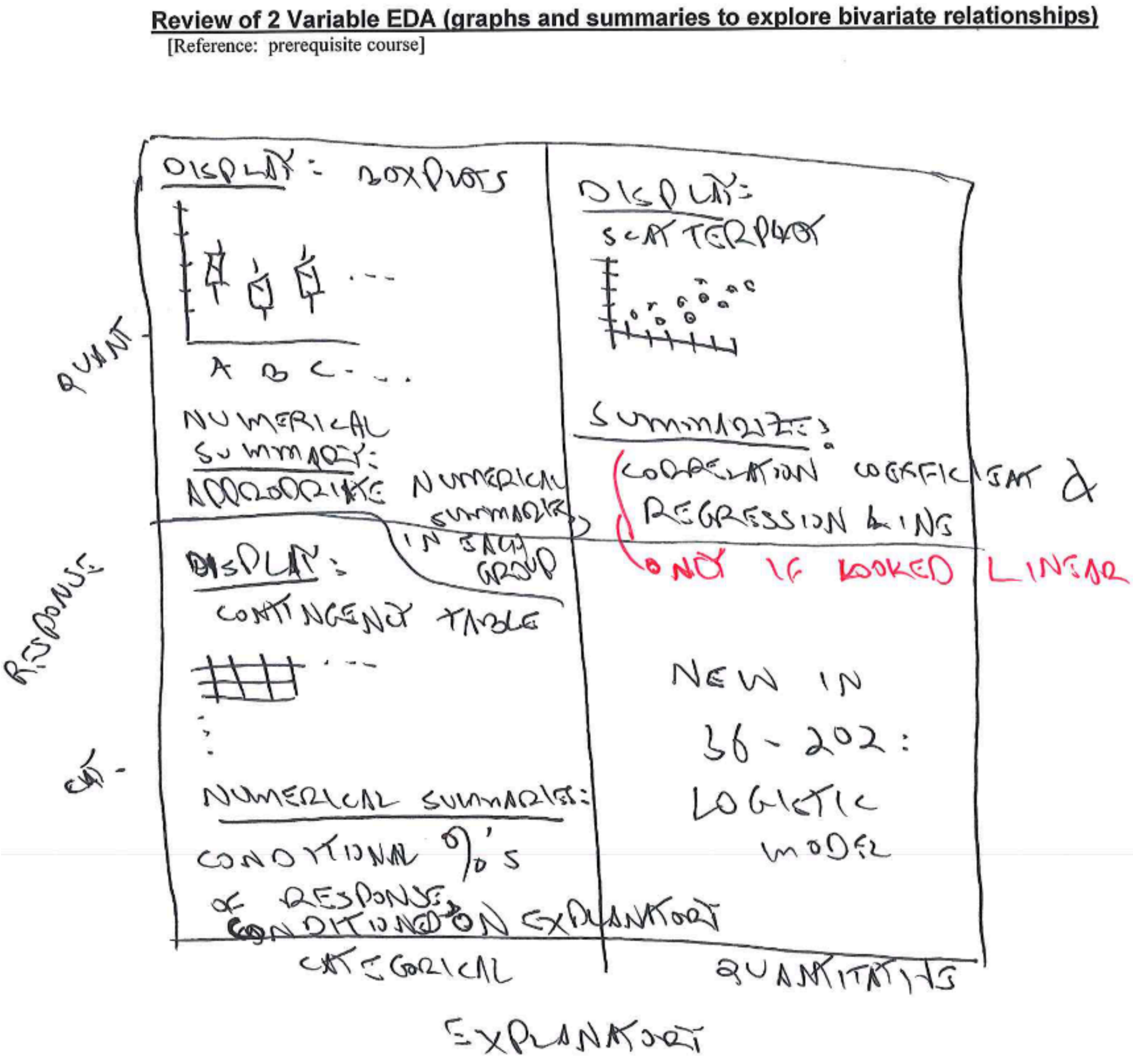
- quantitative
  - histogram
  - center: $\overline{x}$, median
  - spread: stddev, IQR, range
  - five number summary/box plot

1 variable transformations

- need normal distributions?
- $x^{\frac{1}{n}}$, $\log(x + c)$ so everything is > 1.
- the above's inverses
- quantile plots (qqplot) can help us diagnose if normal enough (look for straight line)

2 variable EDA

- explanatory x axis → response y axis



Review of 2 Variable EDA (graphs and summaries to explore bivariate relationships)

[Reference: prerequisite course]

RESPONSE		
QUANT	DISPLAY: BOXPLOTS A B C ... NUMERICAL SUMMARY: APPROPRIATE NUMERICAL SUMMARIES IN EACH GROUP	DISPLAY: SCATTERPLOT SUMMARIES: CORRELATION COEFFICIENT & REGRESSION LINE ONLY IF LOOKED LINEAR
CAT	DISPLAY: CONTINGENCY TABLE NUMERICAL SUMMARIES: CONDITIONAL %'S OF RESPONSE CONDITIONED ON EXPLANATORY	NEW IN 36-202: LOGISTIC MODEL
	CATEGORICAL	QUANTITATIVE
	EXPLANATORY	

1 variable inference

- statistics ($\overline{x}, S_x, ...$) predicts parameters ($\mu, \sigma, ...$)

- components:
  - point estimation: estimate via single number calculated
  - interval estimation: give plausible interview and how plausible
  - significance testing about hypotheses: assess evidence for/against claim about

- 95% confidence interval for μ is $\overline{X} \pm 2 \cdot \mathrm{SE}_{\overline{X}}$
  - (works for arbitrary parameter/statistic estimate)
  - any sample Standard Error SE is $\frac{S}{\sqrt{n}}$ with sample stddev S (but remember, we just use software)
  - technically, 2 should be t_{crit} which varies with n, but it approximates to 2 for 95% confidence when large n

- hypotheses testing
  - H_0 vs H_A
  - "p value is compared to significance level. we do (not) reject the null hypothesis. we do (not) have sufficient evidence that ..."
  - remember: p finds boolean evidence of difference from norm, not magitude of difference

Statistical Model Primer

- statistical models are often of form: quantity $=$ expectation $+$ error
- in 1 variable, eg: $Y_i = \mu + \varepsilon_i$ where μ is the prediction and ε_i is the error at i.
  - we also specify the distribution and mean + stddev of the errors
- in 2 variables, eg: for some X axis value, $Y_i = \mu_{Y|X} + \varepsilon_i$
  - we also specify the shape, center, spread of the distribution of errors

Simple Linear Regression

- our model idea is $Y_i = \beta_0 + \beta_1 X + \varepsilon_i$ where we assume the errors are
  - independent, mean 0, constant stddev/spread (for required for least squares)
  - are normal (required for inference)
  - (can be denoted iid, $N(\mu = 0, \text{variance} = \sigma^2)$)

- our **sample** regression equation is $\hat{y} = b_0 + b_1 X$

- notice that we have three parameters: β_0, β_1, σ
  - they are estimated by b_0, b_1 (when using least squares), and $\hat{\sigma}$: what R calls "Residual standard error"

- to apply the model:
  1. **state** the model
     - eg: "we use the SLR model. $\text{vision distance} = \beta_0 + \beta_1 \cdot \text{age} + \varepsilon_i$ where errors are independent, mean 0, constant stddev, normal.
  2. **validate** the data works for the model
     - linearity: visual inspection
     - errors are:
       - independent: residual plot. residuals "patternlessly" above and below 0 line.
       - mean 0: residual plot. reasonably centered around 0.
       - constant stddev: residual plot. reasonably constant spread, scanning left to right
       - normal: normal qqplot, follows line
     - if there are problems, consider diff model/transformations
  3. **estimate** the parameters
     - use software to find $b_0, b_1, \hat{\sigma}$
  4. **inference**: is data probably showing a relationship between X and Y?
     - t test for $\beta_1 =$ or $\neq 0$
  5. **measure strength** of model with R^2 (if not chance)
     - R^2 is the percent of variability in Y that can be attibuted to the linear relationship with X
     - "Multiple R-squared" in R. NOT "Adjusted"
  6. **predict** of Y from X (for individual with X or all people with X)
     - the equation predicts the point estimate of Y given X
     - get prediction vs confidence interval via R for probable values of Y for individual or all at X

Nonlinear Relationships?

- can use a nonlinear model (same four error assumptions)
- can transform it
- transformations often preferred: fewer parameters make a simpler model
- make sure to not overfit!

Multiple Regression

- we're often interested in predicting a Y from multiple explanatory X_i

- when contribution from each X_i is linear, we have *multiple linear regression*:

$$Y_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon_i$$

  where errors are
  - independent
  - mean 0
  - contant stddev
  - normal

- $p + 2$ parameters: $\beta_{\{0-p\}}$ and σ
  - like SLR, σ is stddev of errors, ie typical deviation of Y from regression hyperplane
  - $\hat{\sigma}$ in R is still "residual standard error"

- each β_i is the avg change in Y when X_i increases by 1 unit and the other Xs remain fixed

- eg `school.mod = lm(GPA ~ IQ + SelfConcept, data=school)`

- to apply the model:
  1. **state** the model
  2. **validate** the data works for the model with EDA
     - scatterplots of Y against *each* explanatory (w/ `pairs` plot). linearity: visual inspection
     - error conditions (also just use a residuals/qqplot):
       - independent: residual plot. residuals "patternlessly" above and below 0 line.
       - mean 0: residual plot. reasonably centered around 0.
       - constant stddev: residual plot. reasonably constant spread, scanning left to right
     - if there are problems, consider diff model/transformations
     - low multicollinearity (each X_i weakly correlated with each other) (might otherwise get mathematically impossible/conceptually inappropriate, misleading results. see `media/high_multicollinearity`)
       - can *informally* investigate via: correlation matrix, odd parameter estimates, oddly large estimate stderrs
       - mathematically diagonse via variance inflation factor (vif)
         - let a model be $Y \sim X_1 + X_2 + X_3$
         - vif of X_i is $\frac{1}{1-R^2}$, with R^2 from $X_i \sim$ the other Xes.
         - i.e., vif of X_1 depends on $X_1 \sim X_2 + X_3$
         - BUT: just use software.
         - when high multicol., drop variables: check diff subsets of Xes, recheck diagnostics for each. find best model with R's *adjusted R-squared* (adjusts for different number of explanatory variables. otherwise, R-squared would be higher with more variables, rmbr?)
         - BUT: also just use software (best subsets routine)
         - vif ≥ 2.5 is concerning
  3. **estimate** parameters w/ software
  4. **inference**: is data probably showing a relationship between X_i and Y?
     - F-statistic: tests if *any* of X_i are important for predicting Y
     - individual T-tests: tests if *each* X_i is a significant predictor *in the presence of all other explanatories*
  5. **predict**: use model, with R^2 for its effectiveness
     - multiple R-squared: proportion of variation in Y that can be explained by all of X_i. has a few properties:
       - closer to 1 = better "fit"

  - can only increase with more predictors
  - diminshing returns

including categorical explanatories?

- check for no interaction between predictors parallel lines iff X_2 doesn't depend on X_1 iff no interaction between X_1 and X_2 iff X_1 effect doesn't depend on X_2.

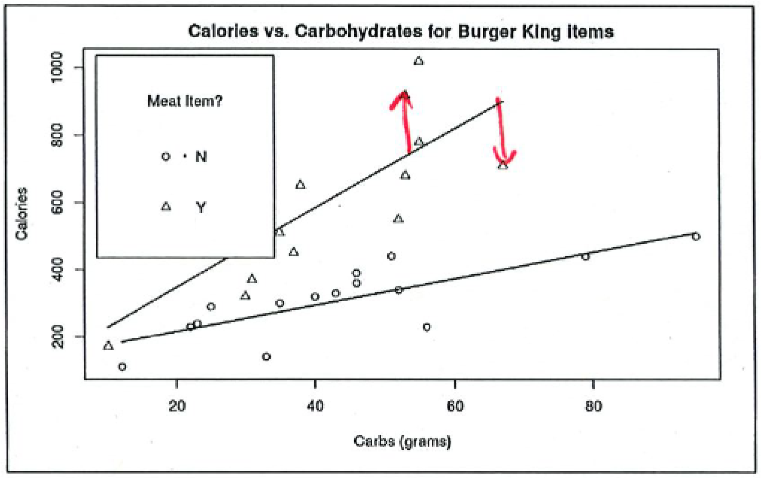
  why is this so verbose from the slides :/

- assuming no interaction between predictors:
  - include a binary indicator/dummy variable (1 if smoker, 0 else)
  - call the category defined as 0 a "baseline" category

- if a categorical variable has, say, 3 options, we get *2* dummy variables, both binary with 0 representing baseline group.
  - "controlling for years of seniority, dept A makes X less than dept C on average"
  - "holding dept constant, we estimate for every extra year of seniority, salary increases by X on average"

what if categorical explanatories *have* interaction?

- let us investigate a situation where calories $\sim$ carbs, but with slopes that differ depending on whether the item is meat.



- new model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot \text{DummyMEAT} + \beta_3 (X_1 \cdot \text{DummyMEAT}) + \varepsilon$$

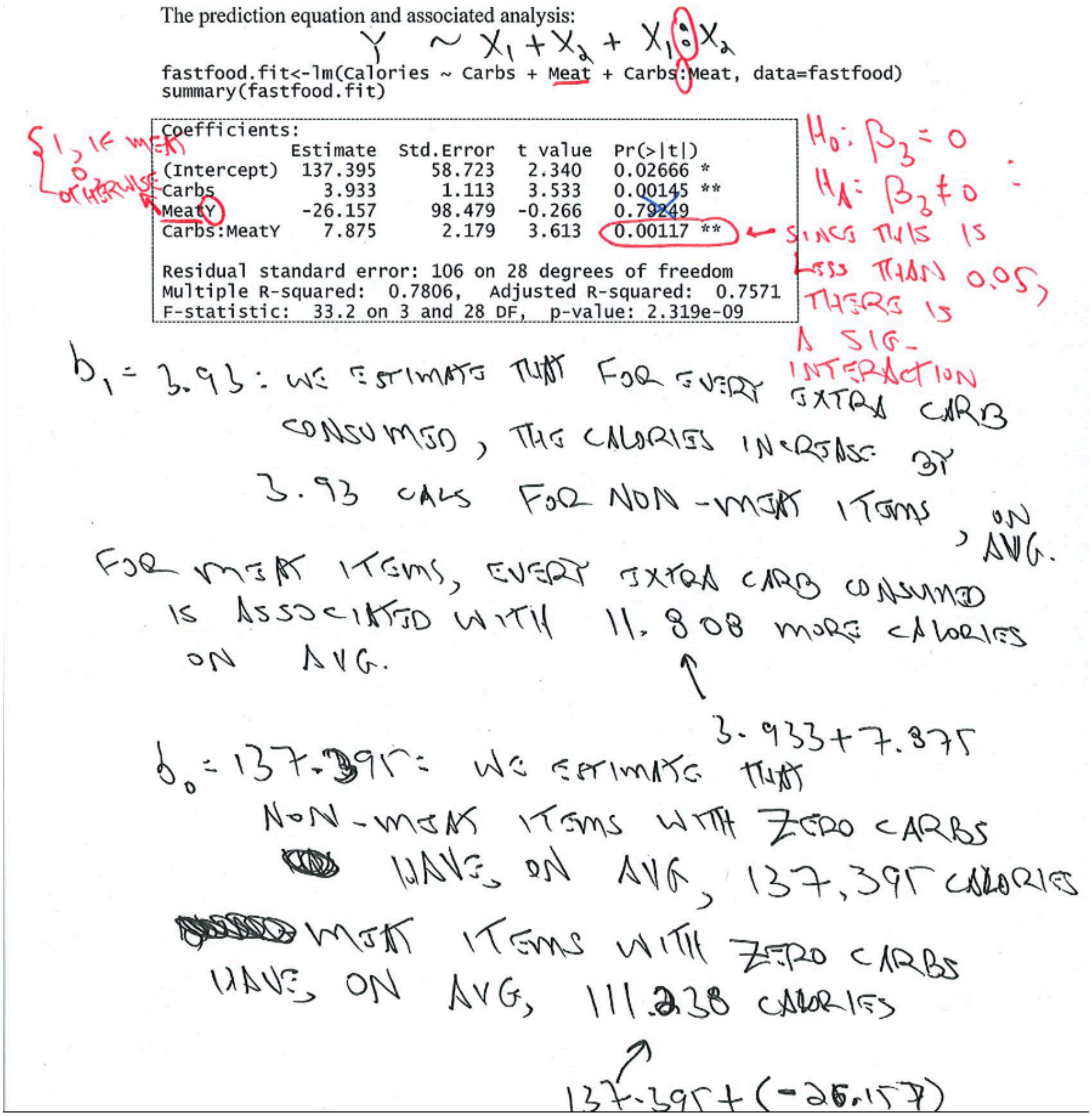
  - capture the difference in slopes with an "interaction term" (the β_3 term above)
  - `lm(Calories ~ Carbs + Meat + Carbs:Meat, data=fastfood)`

- assumptions:
  - population relationship linear within each level of Dummy
  - within each level of Dummy, the errors are indep, mean 0, const stddev, normal (i.i.d., $N(0, \sigma^2)$)

- IMPT: if interaction term stat. significant, then those explanatories must be kept (regardless of their individual variable p-values)

    - this just means that their slopes are indeed different, i think
    - so if they are not significant go back to normal multiple regression ig

- coefficient β_3 interpretation:
    1. difference in slopes; i.e., how the quantitative X_1 effect depends on the group Dummy value
        - “For every unit increase in X_1, the change in Y is β_3 greater/less on avg in Dummy_1 than in Dummy_0
    2. equivalently: how the vertical difference between the lines changes; i.e., how the group Dummy effect depend son the quantitative X_1 value
        - “For a particular value x_1 of the quantitative variable, “



The prediction equation and associated analysis:

$Y \sim X_1 + X_2 + X_1 : X_2$

```
fastfood.fit<-lm(Calories ~ Carbs + Meat + Carbs:Meat, data=fastfood)
summary(fastfood.fit)
```

{1, IF MEAT
0 OTHERWISE

```
Coefficients:
              Estimate  Std.Error  t value  Pr(>|t|)
(Intercept)   137.395     58.723    2.340   0.02666 *
Carbs           3.933      1.113    3.533   0.00145 **
MeatY         -26.157     98.479   -0.266   0.79249
Carbs:MeatY     7.875      2.179    3.613   0.00117 **

Residual standard error: 106 on 28 degrees of freedom
Multiple R-squared:  0.7806,   Adjusted R-squared:  0.7571
F-statistic:  33.2 on 3 and 28 DF,  p-value: 2.319e-09
```

$H_0: \beta_3 = 0$

$H_A: \beta_3 \neq 0$

SINCE THIS IS LESS THAN 0.05, THERE IS A SIG. INTERACTION

$b_1 = 3.93$: WE ESTIMATE THAT FOR EVERY EXTRA CARB CONSUMED, THE CALORIES INCREASE BY 3.93 CALS FOR NON-MEAT ITEMS, ON AVG.

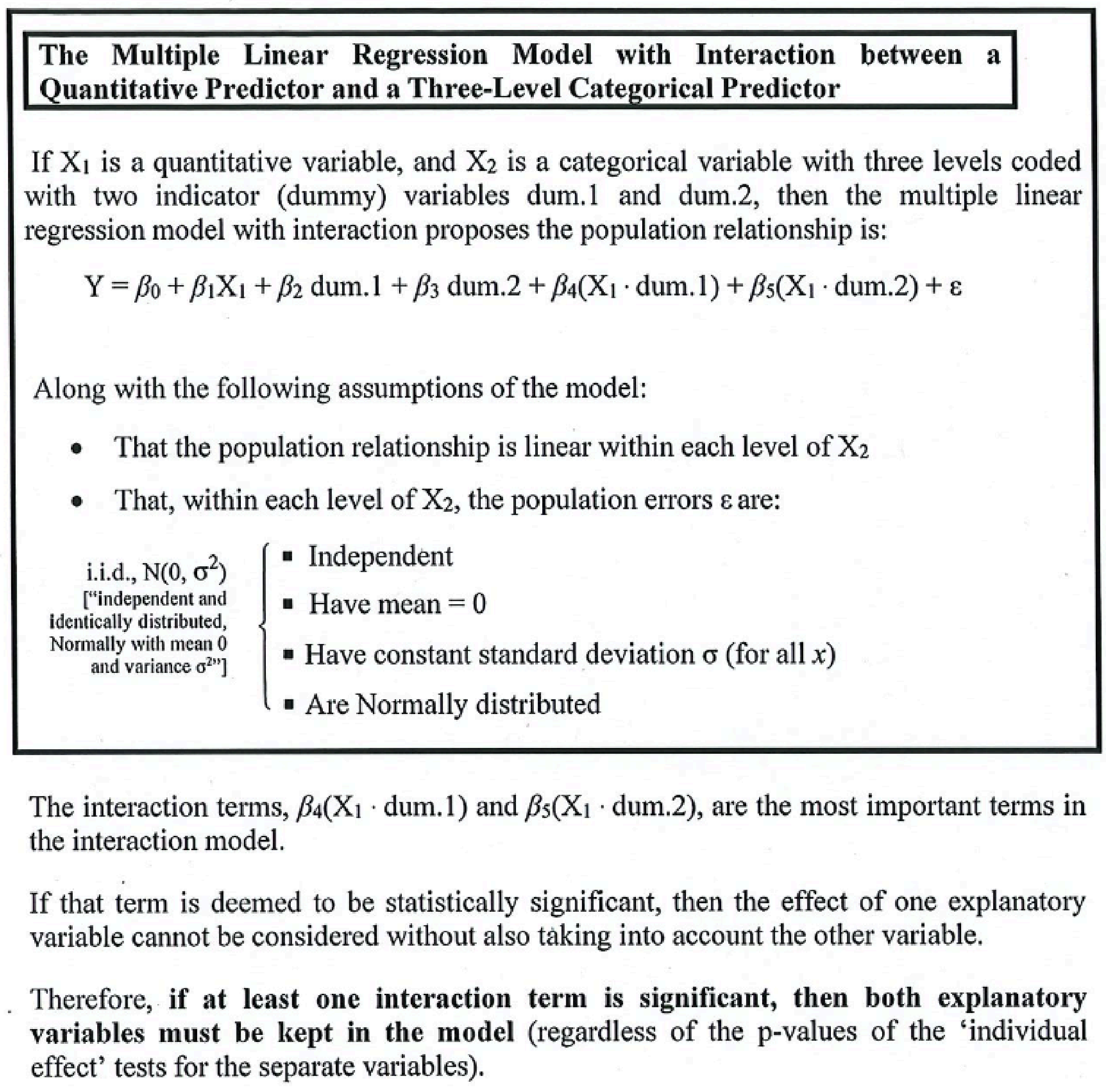
FOR MEAT ITEMS, EVERY EXTRA CARB CONSUMED IS ASSOCIATED WITH 11.808 MORE CALORIES ON AVG.

3.933+7.875

$b_0 = 137.395$: WE ESTIMATE THAT NON-MEAT ITEMS WITH ZERO CARBS HAVE, ON AVG, 137,395 CALORIES

MEAT ITEMS WITH ZERO CARBS HAVE, ON AVG, 111.238 CALORIES

137.395+(-26.157)



The Multiple Linear Regression Model with Interaction between a Quantitative Predictor and a Three-Level Categorical Predictor

If X_1 is a quantitative variable, and X_2 is a categorical variable with three levels coded with two indicator (dummy) variables dum.1 and dum.2, then the multiple linear regression model with interaction proposes the population relationship is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \text{ dum.1} + \beta_3 \text{ dum.2} + \beta_4(X_1 \cdot \text{dum.1}) + \beta_5(X_1 \cdot \text{dum.2}) + \varepsilon$$

Along with the following assumptions of the model:

- That the population relationship is linear within each level of X_2
- That, within each level of X_2, the population errors ε are:

i.i.d., $N(0, \sigma^2)$
["independent and identically distributed, Normally with mean 0 and variance σ^2"]

- Independent
- Have mean = 0
- Have constant standard deviation σ (for all *x*)
- Are Normally distributed

The interaction terms, $\beta_4(X_1 \cdot \text{dum.1})$ and $\beta_5(X_1 \cdot \text{dum.2})$, are the most important terms in the interaction model.

If that term is deemed to be statistically significant, then the effect of one explanatory variable cannot be considered without also taking into account the other variable.

Therefore, **if at least one interaction term is significant, then both explanatory variables must be kept in the model** (regardless of the p-values of the 'individual effect' tests for the separate variables).

good luck on exam 1 <3

One way ANOVA

- recall: we first learn the regression models (Q → Q)
- now we investigate the ANOVA (anal. of variance) models (C → Q)
- big picture: evidence of difference in means of our cat. groups?
  - $H_0 : \mu_1 = \mu_2 = ...$

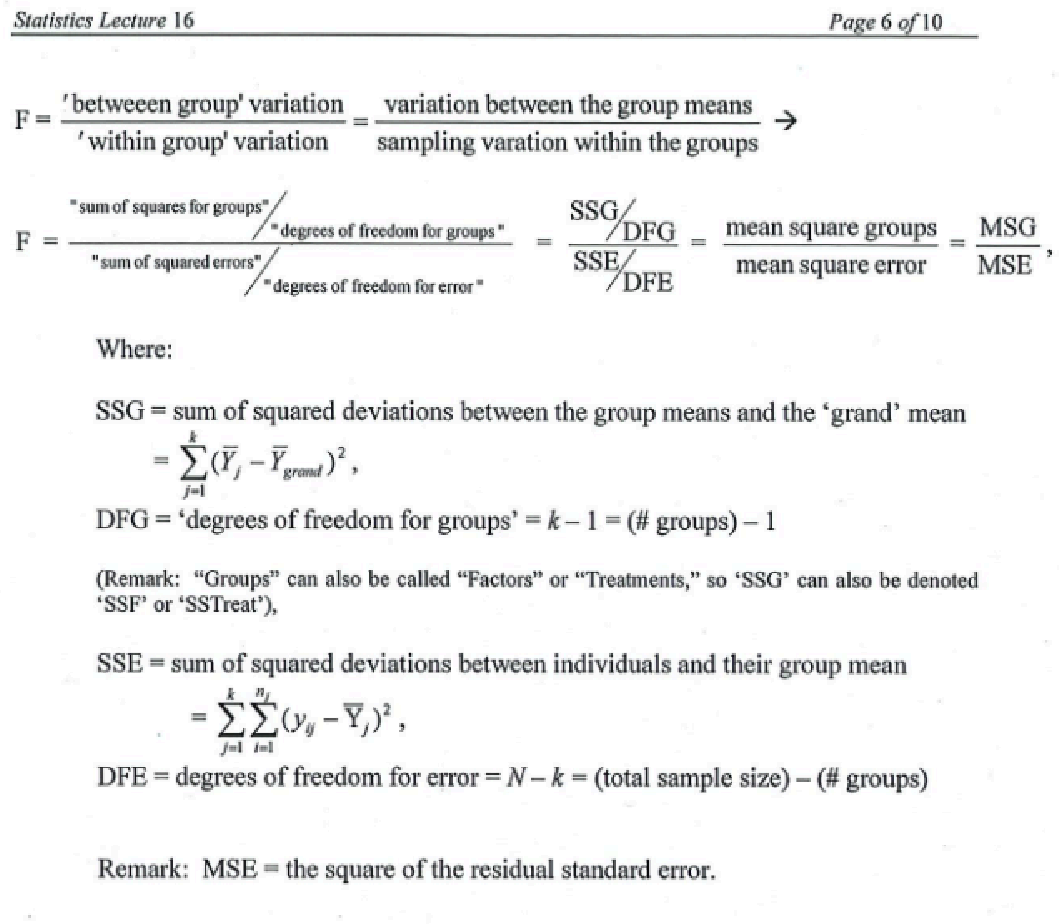
    H_1 : means not all same
- model:

  $$y_{ij} = \mu_j + \varepsilon_{ij}$$

  where j indexes the *independent* cat. populations, i the individuals

  same errors as regression: $i.i.d., N(0, \sigma^2)$

  - parameters: each population mean and stddev of pop. errors
- NOTE: One way ANOVA is stat. identical to mult lin reg with categorical X, $k - 1$ dummy variables
- apply model steps are similar to regression:
  1. **state** the model
  2. **validate** the data works for the model with EDA
     - side by side boxplots (Y vs groups), stats for each gruop
       - if largest stddev ÷ smallest stddev ≥ 2, use $\alpha = 0.025$ in F test, instead of $\alpha = 0.05$
     - groups independent: nature of study. eg no time dependence, 1 person 1 group
     - error conditions:
       - independent: residual plot. residuals "patternlessly" above and below 0 line.
       - mean 0: residual plot. reasonably centered around 0.
       - constant stddev: residual plot. reasonably constant spread, scanning left to right
       - normal: qqplot
     - if there are problems, consider diff model/transformations
  3. **estimate** parameters w/ software
  4. **inference**: is data probably showing a relationship?
     - significant with ANOVA F-test (see above)
     - if yes, suppl. with 'multiple comparisons'
  5. **predict**: use model, with R^2 for its effectiveness
     - multiple R-squared: proportion of variation in Y that can be explained by all of X_i. has a few properties:
- we may derive the ANOVA F-statistic:



Statistics Lecture 16 *Page* 6 *of* 10

$$F = \frac{\text{'betweeen group' variation}}{\text{'within group' variation}} = \frac{\text{variation between the group means}}{\text{sampling varation within the groups}} \rightarrow$$

$$F = \frac{\text{"sum of squares for groups"} \big/ \text{"degrees of freedom for groups"}}{\text{"sum of squared errors"} \big/ \text{"degrees of freedom for error"}} = \frac{\text{SSG}/\text{DFG}}{\text{SSE}/\text{DFE}} = \frac{\text{mean square groups}}{\text{mean square error}} = \frac{\text{MSG}}{\text{MSE}},$$

Where:

SSG = sum of squared deviations between the group means and the ‘grand’ mean

$$= \sum_{j=1}^{k} (\overline{Y}_j - \overline{Y}_{grand})^2,$$

DFG = ‘degrees of freedom for groups’ = $k - 1$ = (# groups) – 1

(Remark: “Groups” can also be called “Factors” or “Treatments,” so ‘SSG’ can also be denoted ‘SSF’ or ‘SSTreat’),

SSE = sum of squared deviations between individuals and their group mean

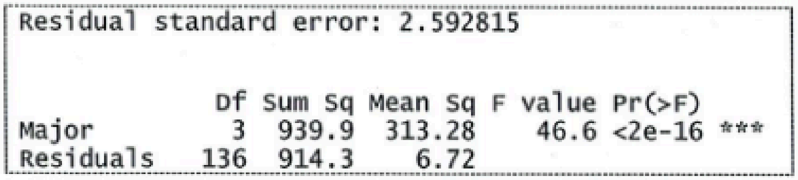
$$= \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \overline{\text{Y}}_j)^2,$$

DFE = degrees of freedom for error = $N - k$ = (total sample size) – (# groups)

Remark: MSE = the square of the residual standard error.

- and ANOVA R^2 value:

Inference, continued: Using R^2 to Evaluate the Strength of the 'X' effect



```
Residual standard error: 2.592815


              Df Sum Sq Mean Sq F value Pr(>F)    
Major          3  939.9  313.28    46.6 <2e-16 ***
Residuals    136  914.3    6.72                   
```

$$R^2 = \frac{939.9}{939.9 + 914.3} = 50.69\%$$

Calculation:

$$\mathbf{R^2} = \frac{\text{sum of squares explained by the model}}{\text{sum of squares total}} = \frac{\text{SSTotal} - \text{SSE}}{\text{SSTotal}} = \frac{939.9}{(939.9 + 914.3)} = 50.69\%,$$

Where:

SSE = sum of squared deviations between individuals and their group mean

$$= \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \overline{Y}_j)^2 ,$$

and in the current example, SSTotal = SSG + SSE, such that:

SSG = sum of squared deviations between the group means and the 'grand' mean

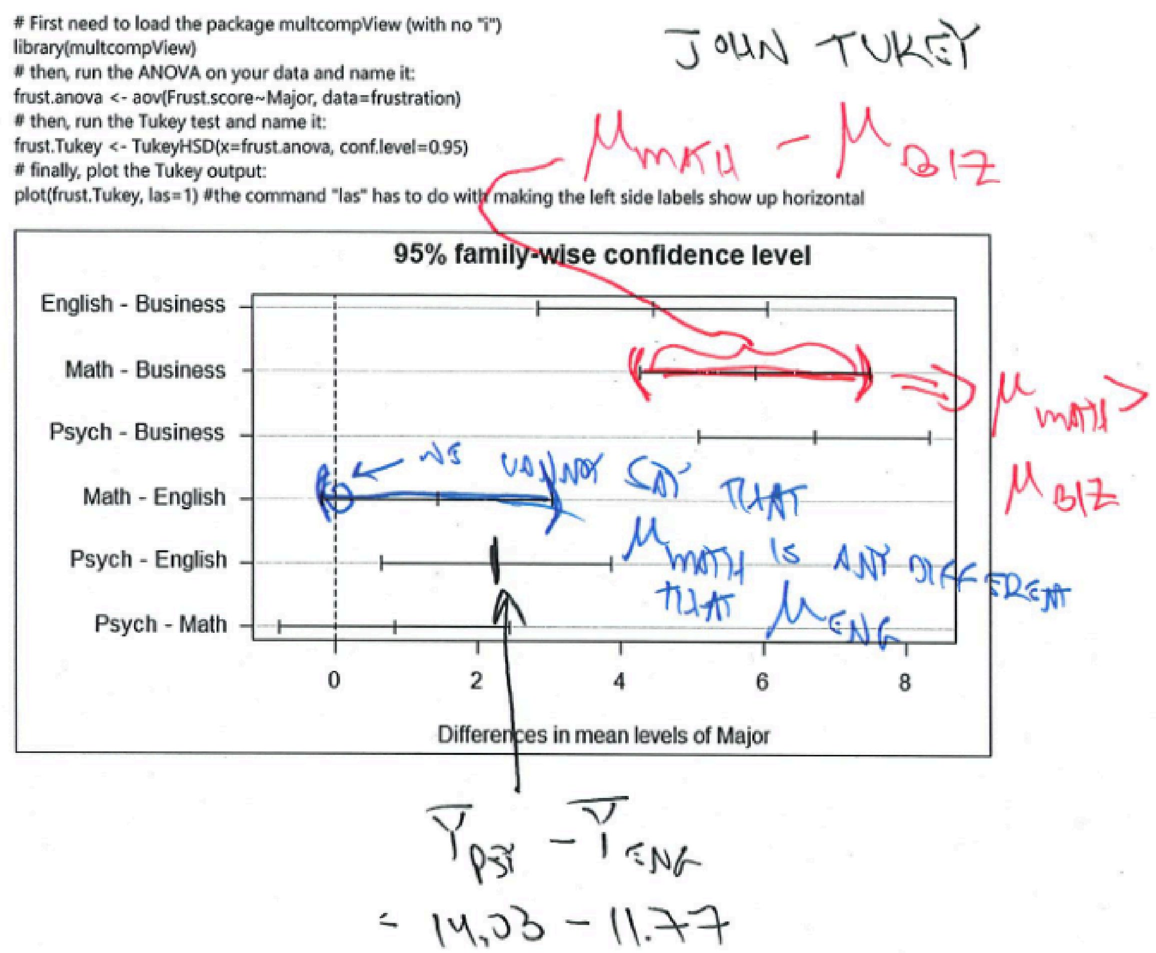
$$= \sum_{j=1}^{k} (\overline{Y}_j - \overline{Y}_{grand})^2 ,$$

(Remark: "Groups" can also be called "Factors" or "Treatments," so 'SSG' can also be denoted 'SSF' or 'SSTreat'),

Interpretation:

R^2 is the proportion of variability in Y attributable to (or predicted from) the explanatory variable(s) based on the model in question. So, in the given example, **50.69% of the total variation in Frustration can be predicted from Major (using the One-Way ANOVA model, with Major as the only predictor).**

- ok, now we believe means not equal. but *which* means?
  - pairwise multiple comparison of means.
  - we can do this via manual CI inspection for means, but more groups mean much higher false positives.
  - so we use Tukey test:



```
# First need to load the package multcompView (with no "i")
library(multcompView)
# then, run the ANOVA on your data and name it:
frust.anova <- aov(Frust.score~Major, data=frustration)
# then, run the Tukey test and name it:
frust.Tukey <- TukeyHSD(x=frust.anova, conf.level=0.95)
# finally, plot the Tukey output:
plot(frust.Tukey, las=1) #the command "las" has to do with making the left side labels show up horizontal
```

JOHN TUKEY

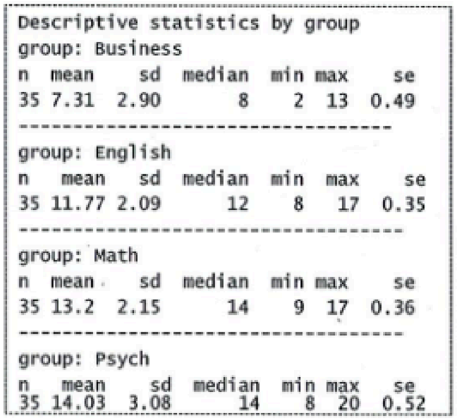
$\mu_{MATH} - \mu_{BIZ}$

$\Rightarrow \mu_{MATH} > \mu_{BIZ}$

VS CANNOT SAY THAT μ_{MATH} IS ANY DIFFERENT THAT μ_{ENG}

$\bar{Y}_{PSY} - \bar{Y}_{ENG}$

$= 14.03 - 11.77$



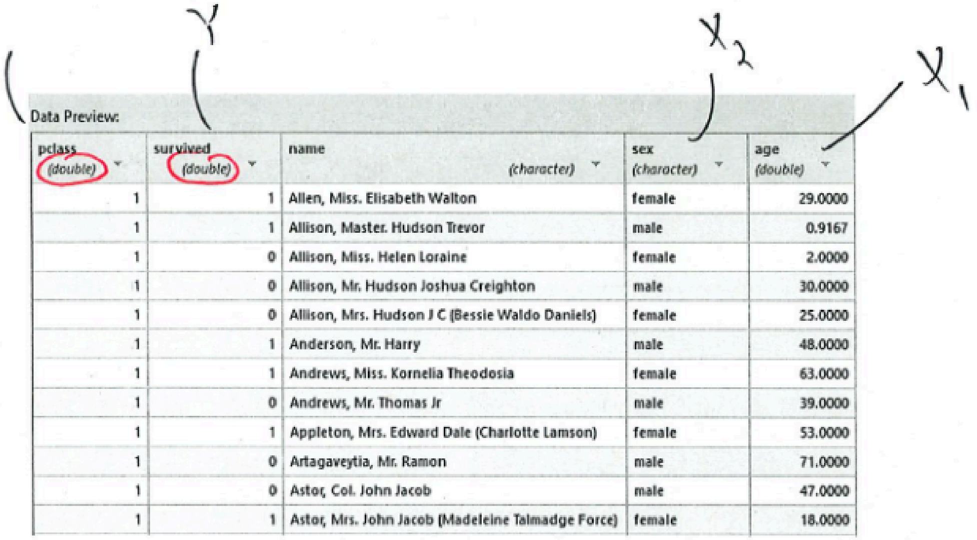
```
Descriptive statistics by group
group: Business
n  mean    sd  median  min max    se
35 7.31  2.90       8    2  13  0.49
-----------------------------------
group: English
n   mean   sd  median  min  max    se
35 11.77 2.09      12    8   17  0.35
-----------------------------------
group: Math
n  mean    sd  median  min max    se
35 13.2  2.15      14    9  17  0.36
-----------------------------------
group: Psych
n   mean    sd  median  min max   se
35 14.03  3.08      14    8  20  0.52
```

Logistic Regression (Simple Binary case)

- simple (one X) binary (Y = 0 or 1)
  - we can get proportion from a dataset eg with `prop.table(table(dataframe$succeeded))`
- logistic regression: Q $\to$ C wow!
- probability $p = P(Y = 1 \mid X = x)$
- suppose Y binary, p defn above.
  - odds favoring Y = 1 are $\frac{p}{1-p}$ (IMPT: this just means if odds=3, odds of Y = 1 are 3:1.)
  - equivalently, $p = \frac{\text{odds}}{1+ \text{odds}}$ via math
- simple binary logistic reg model is

  $$\text{odds}_1 = e^{\beta_0 + \beta_1 X}$$
  $$\text{i.e. } p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

  no error term! $|\beta_1|$ measures "steepness" of log
- linreg: eyeball linearity. hard with logistic. use "goodness of fit" test in R (rg though labs for `ResourceSelection` library, `hoslem.test`)
  - output p value: H_0 : good fit.
- once GOF test passes for sample, test model for appropriateness for population. check β_1 hypotheses test ($H_A : \beta_1 \neq 0$, sig. relationship)
- interpretation:
  - e^{β_0}: odds, on avg, favoring Y = 1 when X = 0.
  - e^{β_1}: for every unit incr in X, odds are *multiplied* by this. (odds ratio!)
- inference: sign. reln? just look at p value for β_1 as usual
  - let's say a CI for β_1 is (lower, upper). to get CI for odds ratio, just e^{lower} etc
- in linreg, R^2 measures association. here, there is no mathy nice measure.
- we use a rough measure: percentage of "concordant pairs"
  - consider: 11 success 14 failures. $11 \cdot 14 = 154$ pairs.
  - a pair is concordant iff the success has a higher prob than the failure in the pair (discordant otherwise)
  - where prob is taken from predicting with the model
- putting it all together:



Data Preview:

pclass (double)	survived (double)	name (character)	sex (character)	age (double)
1	1	Allen, Miss. Elisabeth Walton	female	29.0000
1	1	Allison, Master. Hudson Trevor	male	0.9167
1	0	Allison, Miss. Helen Loraine	female	2.0000
1	0	Allison, Mr. Hudson Joshua Creighton	male	30.0000
1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000
1	1	Anderson, Mr. Harry	male	48.0000
1	1	Andrews, Miss. Kornelia Theodosia	female	63.0000
1	0	Andrews, Mr. Thomas Jr	male	39.0000
1	1	Appleton, Mrs. Edward Dale (Charlotte Lamson)	female	53.0000
1	0	Artagaveytia, Mr. Ramon	male	71.0000
1	0	Astor, Col. John Jacob	male	47.0000
1	1	Astor, Mrs. John Jacob (Madeleine Talmadge Force)	female	18.0000

The Binary Multiple Logistic Regression model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \underbrace{X_1}_{\text{AGE}} + \beta_2 \text{SEX} + \overbrace{\beta_3 \text{dum}1 + \beta_4 \text{dum}2}^{\text{PASSENGER CLASS}}$$

```
titaniclogit <- glm(
  factor(survived) ~ sex + age + factor(pclass),
  family = binomial(link="logit"),
  data = titanic
)
summary(titaniclogit)
```

(Multi)nomial Logistic Regression

Ordinal Logistic Regression

Proportional-Odds Cumulative Logit Model

(end material for exam 2)

Introduction to Statistical Learning

Classifiers

- commonly, we search for cat. response (e.g. pedestrian or not?)
- find boundary in variable space to separate
- EDA: pairs plot

Logistic Regression as Classification

- take log reg, then separate by which side of 0.5 p is on

Discriminant Analysis

- only defined for quantitative X
- assume within each class, variables are Normal
- Linear DA: boundaries are lines/hyperplanes
- Quadratic DA: boundaries are quadratic hypersurfaces
  - assumed Normals can be "stretched" or "twisted"

Classification Trees

- split data into those hyperrectangular things: make decision tree!
- higher nodes are more important
- can use both cat. and quant. X
- CAREFUL overfitting!
  - prune to min decisions manually or via algo
  - split b/t train and test data
  - ensembles
    - bagging trains models and takes avg (RANDOM FOREST!)
    - boosting sequentially combines and error corrects (takes weighted avg)

More about classification tress: The Gini Index

- tree algo decides splits via on purity, often via Gini index
- Gini $\in [0, 1]$ of decreasing purity

Intro to Clustering

- UNSUPERVISED!
- create clusters in variable space

hierarchical aka agglomerative

- start w/ each obs. in a cluster, aggregate based on defn of "closeness"
- "closeness"
  - single linkage: close as nearest nodes (form snakes)
  - complete linkage: close as furthest nodes (form balls)
  - avg linkage: balanced
- assess via adjusted rand index (ARI) $\in [-1, 1]$ of increasing quality

k-means

- hyperspherical clusters in variable space
- very fast algo, simple, better for data w/ Normal around averages
- minimize within-cluster sum of squares (WSS)
- choose k via elbow plot: incr k untill WSS doesn't steeply drop (care overfitting!)