Contents

review

- big picture of applied stats: see 36200 image idk
- we have statistics ($\overline{x}, \hat{p}, ...$) and standard error ($\mathrm{SE}_{\overline{x}}, \mathrm{SE}_{\hat{p}}, ...$)
- 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

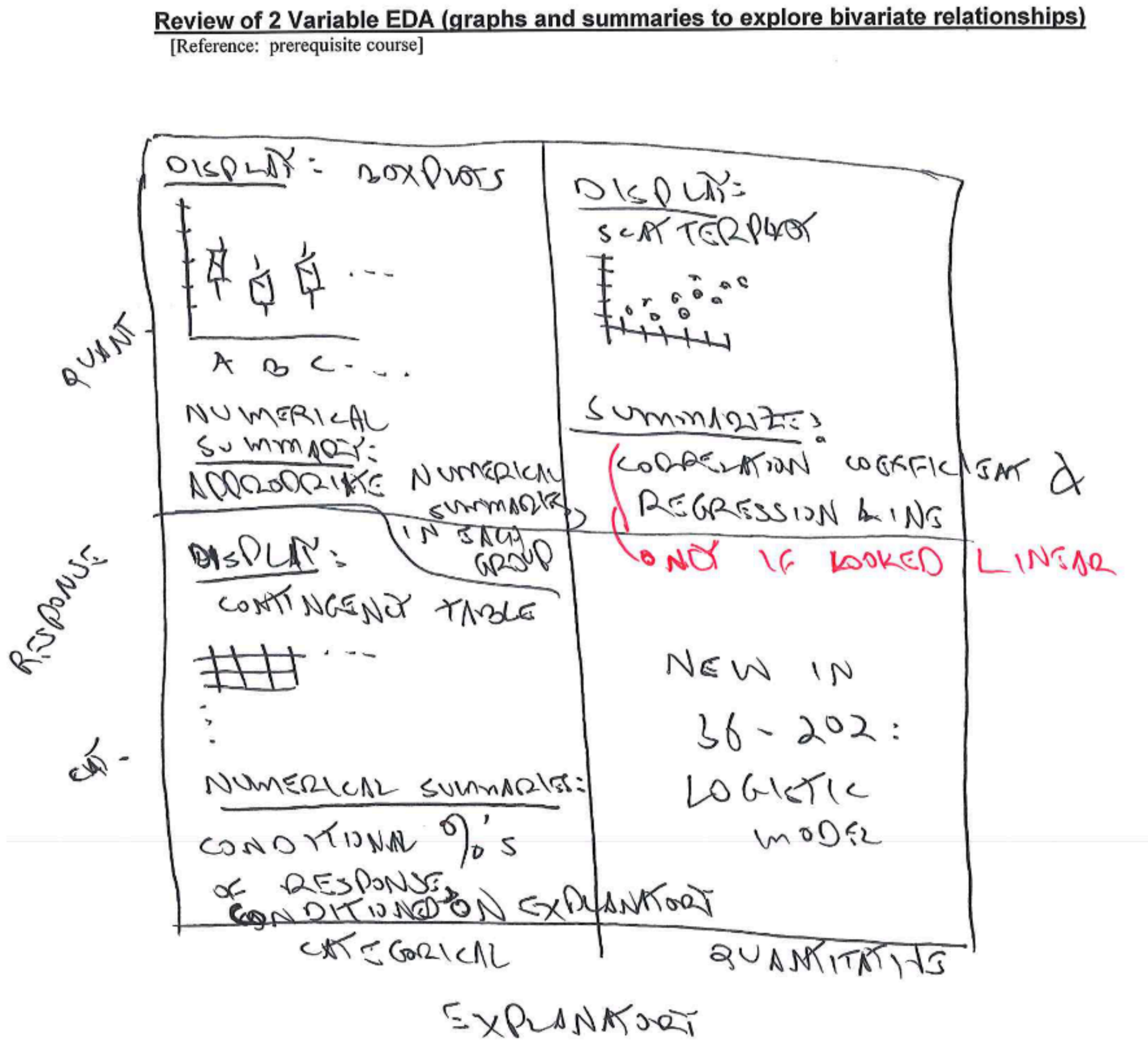
- categorical
  - bar graph
  - percent summaries
- quantitative
  - histogram
  - center: $\overline{x}$, median
  - spread: stddev, IQR, range
  - five number summary/box plot

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 $\rightarrow$ response y axis



1 variable inference

- statistics ($\overline{x}, S_x, \ldots$) predicts parameters ($\mu, \sigma, \ldots$)
- 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 $\overline{X} \pm 2 \cdot \text{SE}_{\overline{X}}$
  - (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
     - 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*:

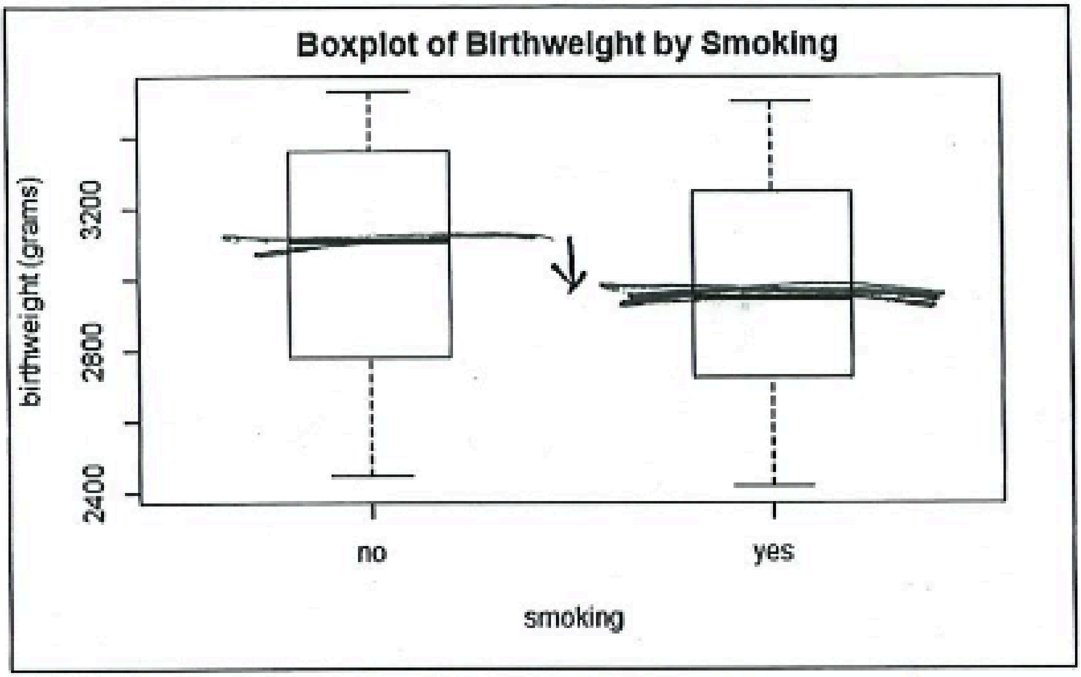
  $$Y_i = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon_i$$

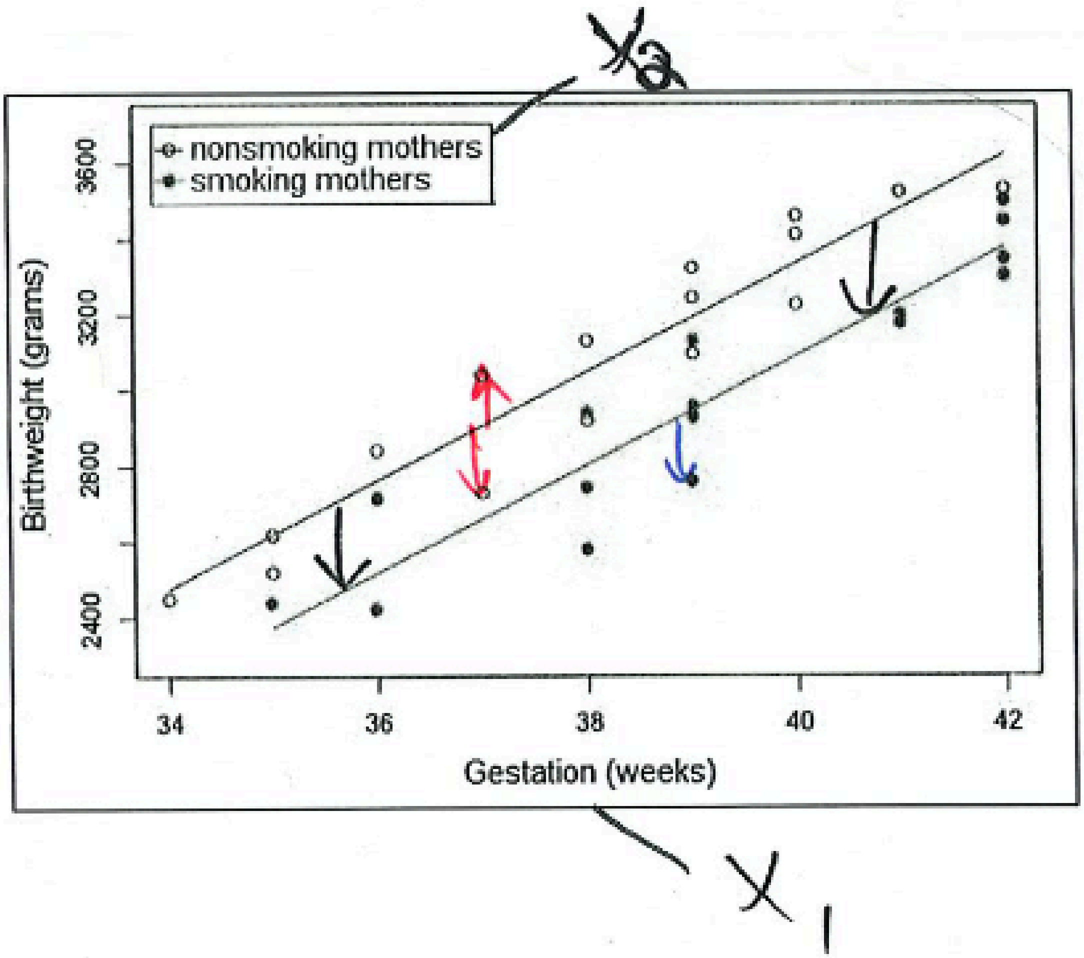
  where errors are
  - independent
  - mean 0
  - contant stddev
  - normal
- $p + 2$ parameters: $\beta_{\{0-p\}}$ and σ
  - like SLR, σ is stddev of errors, ie typical deviation of Y from regression hyperplane
  - $\hat{\sigma}$ in R is still "residual standard error"
- each β_i is the avg change in 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


Boxplot of Birthweight by Smoking
birthweight (grams)
2400
2800
3200
no
yes
smoking


X_2
nonsmoking mothers
smoking mothers
Birthweight (grams)
2400
2800
3200
3600
34
36
38
40
42
Gestation (weeks)
X_1

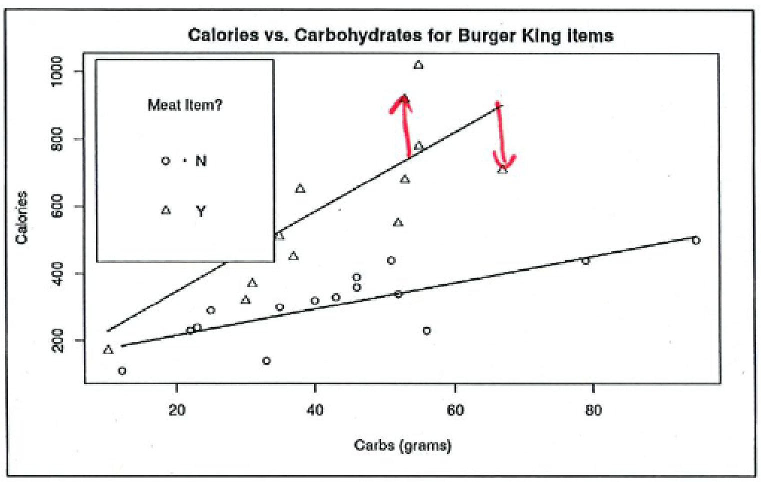
parallel lines iff X_2 doesn't depend on X_1 iff no interaction between X_1 and X_2 iff X_1 effect doesn't depend on X_2.

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 ~ 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, "

ok yknow what tbh just look at this interpretation:

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)
```

$\begin{cases} 1, & \text{if meat} \\ 0, & \text{otherwise} \end{cases}$

```
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)

- and here's a nice lil summary from the notes:

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

If X_1 is a quantitative variable, and X_2 is a categorical variable with two levels, then the multiple linear regression model with interaction proposes the population relationship is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2) + \varepsilon$$

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

good luck on exam 1 <3