~~LECTURE 3: FITTING STATS~~

LECTURE 3: FITTING STATISTICAL MODELS TO DATA IN PYTHON

WHAT DO WE MEAN BY FITTING MODELS TO DATA?

- Note that we're not fitting data to models but models to data.
- Our models must describe the relationships & dist. of the data.

Example: Test Performan & Age

| Variable of interest = Test performance
| Possible predict. = Age

└ Relationship interest: Perf ~ Age. = curvilinear relationship

Goals: (a) Estimate Marginal mean of performance across all ages

(b) Est. Mean performa conditional on age (Relationship of Age w/ Mean Perf)

(I) Model 1 = "Mean only" model for performance

2 Param = Mean & variance. ; Normal dist.

(II) Model 2 = "Conditional" model for performance as a function of age

↳ Normal dist. mean defined by a quad formula & variance (σ^2)

- The data, after examinatn, is normal → Hist, Q-Q plot
- Viz relatnship btw Perform & Age → shows inverse "U" shape relationship

└ scatterplot very useful

(I) "Mean-Only" model = regression model ↦ Regress test perf. on a single mean + error

└ Perf = m + e | errors denote deviation of Obs around the mean

- Check assumptn that errors are Normal. This is done by looking @ residuals via hist. or Q-Q.

(II) "Conditional model

└ Perf = $a + b(Age) + c(Age)^2 + e$ | Params = a, b, and c (regression coef)
e = random error

↓
describe relationship of age w/ performance"

210
180

b ⇒ estimate = linear portion of the equation <funct.>

c estimate → describes acceleration/deceleration in performance as a funct. of age

- Looking @ ratio of estimate to SE; all would seem to be non-zero
- After fitting model; remember to assess fit of model

L Red dashed line = predicted values based on the fitted quad. funct.

Assess fit by looking @ the residuals | In the second plot, there a lot of residuals centered & spread along the predicted mean.

* Let's look @ a poor model. | Predicted values & part show a curvilinear relationship

Note the red dashed lines | our assumption of a linear relationship

WK1

TYPES of VARIABLES IN STATISTICAL MODELING

- There a a dichotomy of data when it comes to dependent variables & independent variables exist in data modeling <model fitting>

1) DEPENDENT variables :→ They are define by our research question

2) INDEPENDENT variables :→ Take are looking and relation with DV & IVs

L Prone to manipulation by an investigator or observed

- Observational studies are mostly focused on describing relationships, ie randomized experiments that enable us to make effect. inferences

ing CONTROL variables in → Randomization helps deal w the confounding

problems in Observational studies

If can't include independent variable included into the model but is a theoretical ~~theorized~~ perspective it helps us control its value when looking @ a relationship where the confounding variable is known

- Missing DATA → We need to find out if the cases that are going to be dropped are systematically diff from those ~~retained~~ retained

L Imputation → predict missing data as a funct. on known data.

WK 1

DIFFERENT STUDY DESIGNS GENERATES DIFFERENT TYPES OF DATA: IMPLICATION FOR MODELING

- Like to table says: different study designs generate different types of data
- It's important to know where the data come from.

Goal = Estimate parameters that best describe the dist of that variable

↙ there is a possibility that different values of a dependent variable of interest maybe correlate. Many times this might be introduced by our study design.

- WK1 OBJECTIVES OF MODEL FITTING: INFERENCE vs PREDICTION

a) Key objectives of model fitting — ⓐ Making inf about rel. betw variables
ⓑ prediction of future outcomes

- OBJECTIVE I → Making inferences
  - Once we have estimate of param. & SE we can → hypothesis test / confi. interval
  - If ~~it~~ is significantly different from 0; then'd be evidence of a linear relationship btw performance & age.
- OBJECTIVE II — Making predictions | regression model = 5-11 + 0.2*(age) - 0.26(age)^2 + e
  - when making prediction from a fitted model; we also need to describe the amount of uncertainty associated w/ the prediction.

WK1 → Reading Notes <BAYESIAN STATS>

- P-values, in frequentist ~~stats~~ statistics, represent the probability of the data (samples) under the null hypothesis. This is not what scientists really want to know - They want to know the opposite: the probability of the hypothesis given the data. This is what Bayesian stats helps us calculate

PLOT

PLOTTING PREDICTIONS & PREDICTION UNCERTAINTY

- Watch out for only looking for point estimates when carrying out predictive modeling. This ain't all!
  - Always plot your uncertainty [the grey band around the line] (u = slope, b = intercept)
    - speaks of uncertainty of the slope
- The second diagram has a lot of uncertainty about its parameters
- 2 ways to account for uncertainty (w/ prediction bounds)
  - (I) Plot data = To look @ variance
  - (II) Look @ standard error = speaks of uncertainty around the prediction estimates

∴ Model 2 needs to worked on w/ caution.

WK 2 LINEAR REGRESSION - INTRODUCTION

Primary outcom (response/variable = Cart Wheel Dist. (inches)) → quantitative

* other variable items (correlation) → Height; Completed Comp

- Data shows reasonable normality. If you were asked to predict the average distance for the next random adult, we could report the mean = 82.48 in
  - But there are other variables that could influence what the CWD at could be
    - So is there a relationship? We examine using a scatterplot
      - When looking @ a scatter plot, we'd like to provide a summary that'll include
        - Form, (as) Direction, (as) Strength, (as) Outliers
          - can be quantified by specifying r & r^2
            - R-squared, the coeff of determination, shows how much of the variability in Y is explained by X. So height only accounts for 11% of the variability in Cart wheel distance

- ~~Our best line of fit~~ ; Estimate of Regression Line ; $\hat{y} = b_0 + b_1 x$

<y = mx + b> - general form

coeff for coeff for
b_0 = intercept, b_1 = slope → Δy over Δx.

└ estimated response when $x = 0$, not usually useful

Ways of Convey us w/the Line of Best Fit

① Least Square Regression = Find the line that minimizes the total squared

<Actual - Predicted> ← Observed / error

└ y intercept; $b_0 = 7.55$; slop; $b_1 = 1.11$

An adult who is 1 in tall than another adult would have a CWD that is 1.1 in longer on average

Predicted CWD = 7.5518 + 1.1076 (Height)

- Extrapolation = any prediction that is outside the range of where ~~remodelled~~ modeled the relationship
- Our regression models gives us an estimated mean response based on the value of x
- Residuals = Observed x value - Predicted x → Used to do some model Checking

WK2 LINEAR REGRESSION INFERENCE

Remember our question: "Is there a significant <+ve linear> relationship btw CWD status and height?"

- A slope of 0 doesn't help because we get same y on diff x
- If we try to plot the entire population on the scatterplot, it'll quickly become busy but we could ~~come up~~ fit the line to the data and come up w/the true regression line

└ There is a true intercept & slope. It is that ~~true~~ underlying (β_1) true slope that we would want to assess i.e. if $\beta_1 > 0$

- Hypothesis tests, H_0: True slope (β_1) = 0, H_A

└ we have a b_1 = 1.1 and we need to find out how far away it is from 0.

- the standard err (0.670) estimates how far away b_1 (estimate) is from β_1 (true value) on average
- the t-statistic (1.653) measures the distance shows how close of b_1 from 0 in standard errs

- so we have a p-value of 0.112 for a two tailed test $\langle H_1 : \beta_1 \neq 0 \rangle$
  - since we are dealing w/ a one tailed test for $H_1 : \beta_1 > 0$,
    the p-value = 0.112/2 = 0.056 (marginally significant @ a 10% level but not @ a 5% level)
- We can also look @ the 95% CI of the β_1 = [-0.278, 2.498]
  (true slope ↙ it covers 0 ↙)

- The confidence bands show the confidence interval bands around the mean response @ x
  - Notice that the bands are curved. The intervals are narrow for values closer to our sample mean (⇒ our $\bar{x}$ reps μ effectively) [Mean response @ a given x]
  - 95% CI for the
    ⇒ because the mean most time rep the true pop mean
- The conf. interval for an individual height a wider than that of the mean.
- Assumption → errors need to have a mean of 0 and a std dev. indep of Y; the height
  $e \sim N(0, \sigma^2)$
- We can't look @ the true errors; only residuals. We use that to check for normality.

Looking @ the Residual Plot — we want to see a random scatter of pts around 0 w/ the points falling in a constant horizontal band, i.e. the variability @ low heights & higher heights are similar.

- Adding another variable; "complete" | Pay attention to the interpretation of the Regression Coefficients

Wk 2 Reading: OVERVIEW OF REGRESSION ANALYSIS + LINEAR MODELS

- The goal of any statistical approach to modeling data is to take a sample of data that represents a population, then use that sample to estimate some facet of the Population.
  - For regression analysis, it's focus is on the CONDITIONAL MEAN VALUE of a single dependent variable (y) correspondly to a given set of predictor variables $(x_1, x_2, \dots x_p)$
- The conditional mean is a function because it depends on the predictor variables.

- Regression analysis = A set of techniques for using a sample of a population represented to then using that data to estimate a conditional mean for that population
  - → Most widely used method = LINEAR LEAST SQUARES
- Linear least squares represent the conditional mean function as a linear function of the predictor variables. <CMF>
- The conditional mean function can be modeled as; $\mathbb{E}[Y|x_1,x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ (population cond. mean; estimate or fit $= b_0 + b_1 x_1 + b_2 x_2$ [b_1, b_0, b_2 est. param])
  - $\mathbb{E}[Y|x_1,x_2]$ read as "the conditional mean of Y given x_1 and x_2"
    - → "expectation" = mean or average value.
  - β_0, β_1 & β_2 = regression parameters → Numeric values that define the CMF (not known & must be estimated from the data)
- Sometimes the intercept is not interpretable. How you would find Age & BMI of 0 to predict the respective Blood Pressure?
  - However, there is a case when the intercept is interpretable, that is when all the predictor variables have been centered to have mean zero. In this case, the intercept = the Cond. mean of the dependent var. when all the predictor var are @ their mean values. The intercept of: Blood Pre. $= \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{BMI})$ would be the average BP for someone who has avg Age & avg height
- Representing a regression model in "generative form"; $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$
  - ϵ = unexplained variation or error or noise
    - represents part of the data that cannot be explained using the predictor variables
- Dependent variable → Response or outcome variable
- Predictor variable → Independent variable, covariate or regressor
- Mean → Expectation, average
- Regression parameter → Slope, coefficient, effect
- Unexplained variation → Error, noise
- Conditional mean function → Regression function
- Constant variance → Homoscedasticity
- Non constant variance → Heteroscedasticity

Linear models →	Linear regression model, linear stat model
Fitted values →	Predicted values
R^2 =	Proportion of explained variance or Coeff. of determination

- The term "linear" in linear least squares is usually used to refer only to (ii) and not to sense (i)

where (i) & (ii) are the senses in which the conditional mean model used in a LLS analysis is linear

(i) - If is linear in the predictor variables.

* (ii) - If is linear in the regression parameters.

Note: It is important to understand that the CMF must be a linear function of the regression parameters if we plan to use linear least squares for estimate

... A NOTE ON CAUSALITY ...

When describing the regression model with CMF say $E[y|x_1,x_2] = 1 + 3x_1 - 2x_2$, avoid saying: "when x_1 goes up by 1 unit and x_2 is held fixed, y goes up by 3 units". Instead say; "when company individuals whose x_1 values differ by 1 unit, and whose x_2 values are the same, the value of y will differ on average by 3 units". → Similarly, instead of saying: "x_1 affects y". it is better to say that "x_1 is associated with y, after controlling for x_2"

- Estimation variance in a regression model is strongly influenced by

3 other characteristics:

(i) the level of conditional variance

(ii) the variance of the predictor variable

(iii) the correlation among the predictor variables

The scatter in the data around the conditional mean. → CV = > EV

This is bad variance because greater CV results in greater uncertainty about the regression parameter

- (i) Refers to how dispersed the values of the diff predictor variables are w/in the dataset being used to fit the model. The more one of a regression model is to establish how diff in the values of a pred. variable relates to diff in the expected val. of the outcome. This is referred to as "good variance" since greater variance in the predictor variable results in less uncertainty about the regression param.

(iii) Correlation among the predictor variables = COLLINEARITY. It makes it difficult for the model to properly attribute a trend to any of the correlated variables.

This doesn't mean that there should be 0 correlation btw predictor variables but we are saying that greater collinearity results in greater uncertainty in the param. estimation. This can be overcome by collecting more data.

Collinearity is viewed as "bad correlation" while correlation btw predictor var. & the dependent var is "good correlation" because they allow us to fit models that do a better job of explaining the variation in the dependent variable.

- EXPLAINED VARIATION / R-SQUARED → The square of the squared Pearson correlation coefficient between the fitted values and the observed value of the dependent variable

— In general, a higher R-squared reflects a better fit of the model but the interpretation should be qualified in 2 ways:

— (i) "Goodness of fit" refers to more than just the mean function & To have a model that fits well, we would like to capture the variance structure as well as the mean structure.

— (ii) Higher R-squared can reflect "overfitting", in which the model fits the data in-hand better than it will fit equivalent data that we observe in the future.

- As noted before, several factors affect parameter estimates $\langle b_0, b_1, b_2 \dots \rangle$;: (i) sample size (ii) conditional variance in the dependent variable (iii) variance of the predictor variable (iv) collinearity

Note that <i & ii> impact all parameters equally ; <iii & iv> impact different parameters in a model to different extents

- CATEGORICAL PREDICTOR VARIABLES

A categorical variable is not numerical and hence can't be directly inserted into the linear predictor : $\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$. This issue is typically addressed by coding the categorical var. into one or more INDICATOR VARIABLES → This can be constructed for each level of a categorical var. It takes 1 for "yes" on a level & "0" for "no" at same level. For example, the indicator var for Gender & Meal eaten (i.e. Breakfast, Lunch, Dinner) = Four. meal of the day → you get ?

	Male	Female	Breakfast	Lunch	Dinner
1	1	0	0	1	0
2	0	1	1	0	0
3	0	1	0	0	1
4	1	0	1	0	0

when including indicator variables as predictor var in a regression model, one indicator derived from the parent categorical variable must be dropped.

- This is because the sum of all indicator var. from the same parent var. is identically equal to 1 and hence the same as the intercept. For the model to be estimable, we cannot include the same variable twice as a predictor variable. Omitting one indicator var. solves this issue and the level that is dropped = THE REFERENCE LEVEL.

- For example, we can drop the "Male" ind. var & only use the "female" indic. var: hence "male" = the reference level.

- The interpretation of a regression param. for an indicator var. should account for the reference level. The regression param. for any non-reference level is the contrast between units w/ that level of the parent variable, and units with the reference level of the parent variable. Example: → The coeff. for the "female" variable would be interpreted as the difference btw females & males. In a model for blood pressure, we might find that the coeff for the 'female' var. = -4. This would mean that holding other factors fixed, females have on average 4 mm/Hg lower blood pressure than males.

- RESIDUALS = the differences between each fitted value from its observed value

A scatter plot of the residuals against the fitted values tells alot about the structure of the fitted model. In particular, if the degree of scatter in this plot INCREASES or DECREASES from left to right, we have discovered a MEAN/VARIANCE relationship

→ This implies that no only is there heteroscedasticity but also the ind. variables differ in a way that is predictable from the mean.

→ Also, knowing that a $\bar{x}/s$ relationship is present indicates that while the Linear Least Squ est. of the regression param are meaningful, their SE may not be correct & we should not rely on the results obtained from standard procedures for stat inf. in linear models

- Scatter plots of the residuals on individual covariates can sometimes reveal nonlinear structure that the model failed to capture. If for example, there is a "U-shaped" relationship btw the residuals and the values of a covariate (e.g x_2), then it better to include x_2^2 or some other transformation of x_2 in the model.

IVQ = In video Question

W2 Logistic Regression - INTRO

N=25

Binary variable of interest = Cartwheel completed

Research Q = Is there relationship betw age & P(cart completed)

- Variable of Int → Takes only 2 values
  - Bulk of data btw 22 & 25
  - Linear regression line doesn't quite fit here.

  → Looking @ scatter plot

- We need to modify our response linear regression so that all our estimates then be between 0 & 1
  - Instead predicting the completion status using a linear regression model, we'll predict a transformed version of the probability of success
    - → LOGIT TRANSFORMATION (uses the Logit function or Log odds)

Logit function = $\ln\left(\frac{p}{1-p}\right)$; p = Probability of success

→ odds = P(success) / P(Failure)

- The Logistic function of RHS = A symmetric distribution

NOTE: The higher the value of the logit function, the higher the prob. of success (vice versa)

- There are no maximums of the logit function - They extend to $-\infty$ & ∞

Our model • $\text{logit}(\hat{y}) = b_0 + b_1 x$

Scatter plot + Logistic Regression line → The logistic regression line is a curved line because we're looking @ the P(success) scale

→ Don't extrapolate i.e. estimate y based on an x that is not within the range of the data.

Python output:

	Coef
Intercept	-4.4213
Age	0.2096

; $\text{logit}(\hat{y}) = -4.42 + 0.2096(\text{Age})$

- Easy attempts to interpret
  - 1.23 = $e^{0.2096}$

- Reduction Uncertainty = Bands are smaller where near the mean $\bar{x}$.
  but where there are fewer observations, the bands are wider.

- Transform log-odds to probabilities $\langle \hat{y} = \hat{p}(\text{Compl. Carbon}) \rangle$ → The predicted curves
  Mainly because our n is small ← bands are curved but not smooth

Assumptions = ① The model; logit($\hat{y}$) has to be linearly associated with dependent variable
② Residual plots aren't that helpful since $\hat{y}$ takes only 2 values. This can be solved by by having more variation & also having more covariates

WK 2 LOGISTIC REGRESSION LINE

- Confidence Interval $\langle 95\% \rangle$ for Slope: Sample slope ± "a few" · Est. SE'

From python results: $0.2096 \pm 1.96(0.171)$ ← $b_1 \pm z^* SE_{b_1}$

95% CI = $(-0.126, 0.545)$ ⇒ NOT significant because contains 0

- Hypothesis testing: $H_0: \beta_1 = 0$; $H_A: \beta_1 \neq 0$

$$\text{test stat} = \frac{b_1 - 0}{SE_{b_1}} = \frac{0.2096}{0.171} = 1.225 \rightsquigarrow p\text{ value} = 0.221$$

WK 2 Reading: OVERVIEW of REGRESSION w/ BINARY OUTCOMES

- The linear approach is useful for many diff types of variables; especially quantitative variables. However, often the need arises to conduct a regression analysis using a binary dependent variable.

↳ Ex: Consider a research study on the US adult population and look @ whether a given person in this population has ever been a smoker in their lifetime. This is a categorical variable w/ 2 levels, ie it doesn't have an expected mean [i.e. the mean of the variable values, if coded as "smokers" & "non-smokers", means nothing]. However we can choose one cate, say the "smoker" cate. and aim to model the probability that this outcome occurs. Equivalently, we can code the outcome numerically as 1 (for "smoker") and 0 (="non-smoker") and then model the expected value of this binary equivalent. [both approaches are equivalent]

A binary dependent variable such as smoking status generally can't be modelled with linear regression because trying to model a binary dependent variable w/ linear regression

would often obtain fitted probabilities that are greater than 1 or less than 0.

Link Function → used when working w/ binary dependent variables and is used to map the probability (or mean) to a value on the real no. line where can then be modeled using the linear predictor. Logistic function or Log odds function ($\log\left[\frac{p}{1-p}\right]$) is the link function commonly used for regression w/ a binary outcome

- **Odds** → A quantity derived from a probability. It can take on any non-negative number

$L = \frac{p}{1-p}$. The odds of 3 = we are 3 times more likely to observe a 1 than to observe a 0

- The odds of $1/3$ = we are 3 times more likely to observe a 0 than to observe a 1.
- The odds = 1 is neutral because we are equally likely to observe both outcomes. $\left(\frac{1/2}{1/2}\right)$
- A further transformation that is often applied is to take the (natural) logarithm of the odds yielding the log odds - $\log\left(\frac{p}{1-p}\right)$. The neutral pt for the log odds = 0 ; i.e. $\log(1) \Rightarrow p = 1/2$

The log odds is symmetric around 0 in the sense that a log odds of v and a log odds of $-v$ convey the same strength of association in terms of the base relative to the neutral value. This symmetry makes log odds a good choice for modeling via the linear predictor

Logistic Regression

Here, we set $\log\left[\frac{p}{1-p}\right] = b_0 + b_1 x_1 + \ldots + b_p x_p$. For example, if the outcome we are modeling is smoking history, and the covariates are age and gender (coded as female = 1 if the ref. level of this categorical var. is "male") then the model is

$$\log\left[\frac{p}{1-p}\right] = b_0 + b_1(\text{age}) + b_2(\text{female})$$

Company w/ subpop of people have a food gender

— If b_1 is +ve, then older people are more likely to be smokers than younger people; whereas if b_1 is -ve, older people are less likely to be smokers than younger people.

If $b_1 = 0$, then age is unrelated to the probability that a person smokes (w/in a specific gender)

— If b_2 is +ve, then females are more likely to smoke than males and if b_2 is -ve, then females are less likely to smoke than males.

PARAMETER INTERPRETATION

The quantitative interpretation of the logistic regression param. are best expressed;
- (i) additively in terms of the log odds or
- (ii) multiplicatively in terms of the odds.

Example: Age parameter 0.04 ; same gender comparison but ages differ by one year

(i) The older person has log odds for smoking that is 0.04 units greater than that of the younger person. This relationship is linear in age, so if we compare two people of the same gender whose ages differ by 10 yrs, then the older person has 0.4 greater log odds for smoking than the younger person.

(ii) Note that when exponentiating the log odds to get the odds they are no longer linear but rather they're multiplicative. Since $\exp(0.04) \approx 1.04$, comparing two people of the same gender whose ages differ by one year, the older person has 1.04 greater odds of smoking than the younger person. When comparing people of the same gender whose ages differ by 10 yrs, the older person has $\exp(0.4) \approx 1.49$ times greater odds of smoking than the younger person.

WK 2 PROCEDURES → MAY CHANGES DEPENDENT

R-Squared can be defined as — (i) Squared Pearson corr. coef between the covariate & the outcome for just 1 outcome & 1 covariate ← (multiple covariates) (ii) Squared Pearson corr. coef between the fitted values & the outcome

WK3 WHAT ARE MULTILEVEL MODELS & WHY DO WE FIT THEM?

- This week we'll be looking @ appropriate stat. models for datasets that are generated by study design that introduce dependencies in the data.
- Now we're working with data where the variables are correlated

Multilevel models = used to model dependent data

that makes it unlike — here, the regression coeff vary <& not fixed> across the higher level clusters.

helps estimate the variability amongst subjects & clusters in terms of the coeffs of interest

- We estimate relationships & parameters that estimate variances

Level 1 : $Y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + e_{ij}$

Dependent variable, y, which is defined for a given observation (i), i = obsv, j = cluster, nested within cluster (j).

∴ these coeffs are refered to as random coeffs and not params because they change w/ j

* Notice how the Y is a function of coeff which have a subscript (j) which means that the regression coeff are determined by what cluster, j, we are refering to. They randomly vary, depending on the cluster.

Level 2: $\beta_{0j} = \beta_0 + u_{0j}$ → the random effect for each cluster that makes the random coeffs vary.

$\beta_{1j} = \beta_1 + u_{1j}$ → the regression parameter

Below level 1 & there are eqns for the random coeffs @ level 1

• Multilevel models allow us to decompose the unexplained variance in a given outcome into (i) between- & (ii) within-cluster variance

*Random effects capture (i) while error terms @ level 1 capture (ii)

Key Question: How much of unexplained variance due to between-cluster variance arises in the intercepts or slopes of a given model?

WK3 **MULTILEVEL LINEAR REGRESSION MODELS**

Model for a continuous dependent variable Y_{ij} measured on person i within cluster j

$$y_{ij} = \underbrace{\beta_0 + \beta_1 x_{1ij}}_{\text{Fixed effects}} + \underbrace{u_{0j} + u_{1j} x_{1ij}}_{\text{Random effect}} + e_{ij}$$ ← error term: captures what's not predicted by the regression coeffs & the rand. effects

Fixed param that capture the overall intercept in the model & overall relationship of x with y

Random effect: capture btw cluster variance

- Fixed effects are used to find Mean of the dependent variable & error
- Random effect: require a defined distribution ie $[e_{ij} \sim N(0, \sigma^2)]$

EBLUPS = Empirical Best Linear Unbiased Predictions.

WK3 **MULTILEVEL LOGISTIC REGRESSION MODELS**

$$\ln\left[\frac{P(y_{ij}=1)}{1-P(y_{ij}=1)}\right] = \text{logit}\left[P(y_{ij}=1)\right] = \underbrace{\beta_0 + \beta_1 x_{ij}}_{\text{fixed effects}} + \underbrace{u_{0j} + u_{1j} x_{1ij}}_{\text{Random effects}}$$

single level equation incorporates rand. effects

Q: Suppose that you examine the variability among randomly sampled higher-level clusters of observations (e.g. schools) in the proportions on a binary variable of interest, and you find visual evidence of significant variability in the proportions. You then fit a logistic regression model to these data in Python, but forget to fit a multilevel model including random effects of the clusters; how would this affect your analysis?

* By omitting explicit random effects in the logistic regression model, the standard errors of our estimated fixed effects would likely be too low, and the estimates of the fixed effects would likely be incorrect because we're failing to explicitly adjust for the random effects of the higher level clusters. Omitting random effects when they are important is the same type of model specification error as omitting the fixed effect of an important predictor variable.

WK3 **WHAT ARE MARGINAL MODELS & WHY DO WE FIT THEM?**

- We don't include random effects in marginal models because we don't care about the Between-cluster variance rather we want to model the WITHIN cluster variance.
- If we fail to account for the fact that observations are correlated w/in the same higher level cluster, we run the risk of UNDERSTATING our standard errors.

WK4 **SHOULD WE USE SURVEY WEIGHTS WHEN FITTING MODELS**

WK4 **BAYESIAN APPROACHES TO STATISTICS & MODELING**

- Bayesian statistics @ belief & probability