General Linear Models with One Predictor

- Topics:
  - Vocabulary and broad categories of predictive linear models
  - Special case of GLM 1:
    "(Simple) Linear Regression" with a quantitative predictor
  - Special case of GLM 2:
    "Independent (or two-sample) t-test" with a binary predictor
  - Foreshadowing uses of the GLM

Review: Methods to Answer Univariate and Bivariate Questions

- **Univariate mean comparisons** and what they are known as:
  - "**One-sample z-test**": Used to test a sample mean against an expected population mean (the H_0) using a known variance (and big enough N)
  - "**One-sample t-test**": Used to test a sample mean against an expected population mean (the H_0) using an unknown variance—because variance must be estimated, we need to correct for denominator DF remaining (N)
- **Bivariate association indices** for different types of variables:
  - "**Pearson's r**": Used to quantify linear relationship between two quantitative variables; r is tested for significance against H_0 (e.g., 0) using t-distribution
  - "**Spearman's rho**": Pearson's r using rank-ordered versions of quantitative variables instead, which is more appropriate for quantitative variables with concerning extreme values or for ordinal variables (i.e., numbers are labels)
  - "**Pearson's χ^2**": Test if association between categorical variables is $\neq 0$ using χ^2 distribution; χ^2 must be converted to an effect size (e.g., r, risk ratio, odds ratio) to quantify strength of association independent of significance

Steps in Quantitative Data Analysis

- **Quantitative data analysis**: the process of applying statistical models to a sample of data to answer your research questions
  - Enter, download, or otherwise acquire quantitative data
  - Import data into statistical software and verify its accuracy
  - Describe data using univariate statistics and bivariate measures of association; use these to double-check accuracy of data

- **Select a family of statistical models** based on the characteristics of the variables of interest and the questions to be answered
  - Estimate statistical models, check results for potential problems…
  - Estimate more statistical models, check results again…
  - Estimate even more statistical models… interpret results!
  - Write up the results: Btw: you did not "run analyses" or "calculate models"; you "conducted analyses" and "estimated models"

Roles and Labels of Study Variables

When research questions are phrased as *what is the role of x in explaining y*, below are possible synonyms of x and y:

- <u>Reason (Explainer):</u>
  - In notation: x variable
    - Exogenous (is not explained)
  - **Predictor**
    - My preferred generic term
  - Independent variable (**IV**)
    - Used more often when variable is manipulated (like treatment)
  - Covariate
    - Used for reasons the researcher is not interested in (but must include to keep others happy); also used for quantitative predictor in ANCOVA

- <u>What is To Be Explained:</u>
  - In notation: y variable
    - Endogenous (is explained)
  - **Outcome**
    - My preferred generic term
  - Dependent variable (**DV**)
    - Used more often in experimental studies
  - Criterion
    - Used in observational studies with "regression" models

Roles of Variables: Some Examples

- In the following example research questions, identify which variables are **predictors or outcomes** and their likely types:
  - To what extent does positive feedback improve performance speed and accuracy more than neutral feedback?
    - Predictors:
    - Outcomes:
  - Is faster academic growth in elementary school related to more frequent reading to children when in preschool?
    - Predictors:
    - Outcomes:
  - How effective is teacher training for creating higher rates of positive feedback to a teacher's students?
    - Predictors:
    - Outcomes:

Types of Inferences: 2 possibilities in describing how x relates to y

- **x causes y → causal inference** requires the following:
  - x variable had to come first (temporal precedence)
  - x variable was under complete experimental control during the study (i.e., through random assignment and experimental manipulation)
  - Study design eliminates all possible alternative explanations
- **x relates to y** (synonyms = **associative**, **correlational**)
  - We have observed a relationship, but we do not have the ability to infer cause given the design (i.e., it's an observational study without control)
  - In lieu of experimental control, we can attempt **statistical control**: include other predictors that represent alternative explanations for why x relates to y, and see if x is still related to y → many research questions try to do this
- These 2 possibilities can only be distinguished by study design—they have nothing to with the type of variables collected (a common misconception)
- Because causal inference is rarely possible in studies of real people, we will **only use associative language** in describing model effects in this class

Moving On to Predictive Linear Models

- Questions concerning more than variables at a time are best answered using **predictive linear models**, in which one must designate which variables are predictors and which are outcomes
- Models come in different flavors based on type of outcome variable
  - Continu-ish quantitative outcome?
    - "**General**" Linear Models using the normal distribution—us this semester
  - Literally any other kind of outcome variable?
    - "**Genera*lized***" Linear Models using some other distribution and a transformed predicted outcome (called a "link function") to address variable possible values and boundaries—here are some examples:
      - Binary outcome? Use Bernoulli distribution and logit link
      - Ordinal outcome? Use multinomial distribution and cumulative logit link
      - Nominal outcome? Use multinomial distribution and baseline logit link
      - Binomial outcome? Use binomial distribution and logit link
      - Count outcome? Use Poisson-family distributions and log link
    - Come back in Spring 2022 to learn these genera*lized* linear models ☺

What "Linear" in "Linear Models" Means

- Most predictive models have a "**linear**" form, which looks like this:
  - $y_i = (\text{constant} * 1) + (\text{constant} * \text{Xpred1}_i) + (\text{constant} * \text{Xpred2}_i)\ldots$
  - Fortunately, this does NOT mean that we can ONLY predict linear relationships—we can specify many nonlinear forms of relationships of quantitative predictors (the Xpred_i variables) as needed or expected
  - Fortunately, this also means we can include categorical x_i predictors

- Historically, variants of the **general linear model** (for continu-ish outcomes) get siloed into different classes and called different names based on **what kind of x_i predictor variables are included**:
  - Called "(Linear) (Multiple) **Regression**" if using quantitative predictors
  - Called "Analysis of Variance" (**ANOVA**) if using categorical predictors
  - Called "Analysis of Covariance" (**ANCOVA**) if using both predictor kinds
  - We are going to cover all of these as special cases of the General Linear Model ("**the GLM**")—separating them does way more harm than good
    - We will use SAS GLM (REG for standardized) and STATA regress for all!

Welcome to the GLM!

- Linear models use **new notation within one equation** to describe how all the x_i predictors relate to the y_i outcome(s) in your sample
  - **1** outcome? "**Univariate** GLM" **2+** outcomes? "**Multivariate** GLM"
- Starting point for univariate GLMs is always to represent central tendency and dispersion of the outcome variable (y_i)
  - We will use mean and variance to describe the outcome because the GLM uses the normal distribution (in which skewness should be 0)

- Your first GLM is the "**Empty**" model (=no predictors): $\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{e_i}$
  - $\boldsymbol{y_i}$ = "y sub i": outcome variable for *each person* in your sample
  - $\boldsymbol{\beta_0}$ = "**beta 0**" (sometimes called "beta not" but not by me)
    - More generally, betas ($\boldsymbol{\beta}$) will represent **values to be estimated** that will apply to the whole sample (i.e., betas are constants) = "**fixed effects**"
    - The beta **subscripts index each fixed effect** (starting at 0)

The "Empty" General Linear Model

- The "**Empty**" model (empty = no predictors): $\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{e_i}$
  - $\boldsymbol{\beta_0}$ = "beta 0" = "**the intercept**" (or "the constant", ugh) and is defined as the predicted (expected) value for the y_i outcome when all x_i predictors = 0 (so the estimated value for $\boldsymbol{\beta_0}$ will change as the predictors are changed)
  - We don't have any predictors yet, so the intercept takes on the single most likely value for everyone—the **sample** (or "**grand**") **mean** (so in this model, $\boldsymbol{\beta_0} = \bar{\boldsymbol{y}}$)
- So what would $\boldsymbol{\beta_0}$ be for:
  - The blue line? the red line?
  - But why do the red and blue lines differ????

Univariate Normal PDF:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} * \exp\left[-\frac{1}{2} * \frac{(y_i - \mu)^2}{\sigma^2}\right]$$

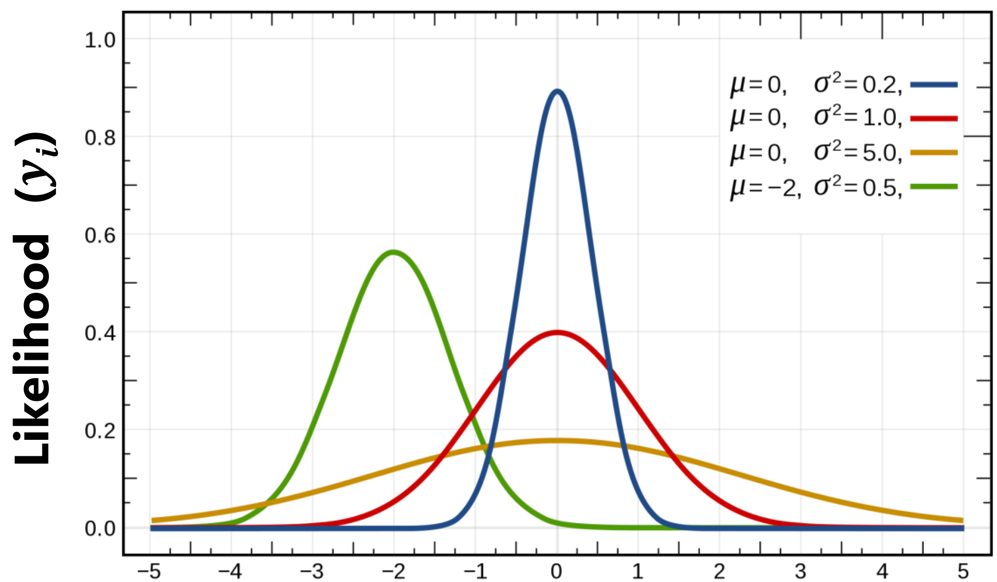


Image borrowed from: https://en.wikipedia.org/wiki/Normal_distribution

The "Empty" General Linear Model

- The "**Empty**" model ("no predictors"): $\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{e_i}$ (in which $\boldsymbol{\beta_0} = \boldsymbol{\bar{y}}$)
  - $\boldsymbol{e_i}$ = "e sub i" or "**residual**" = deviation between the actual $\boldsymbol{y_i}$ outcome for each person and $\boldsymbol{y_i}$ outcome predicted by the model (through the beta fixed effects)
  - Because the empty model predicts the same $\boldsymbol{\bar{y}}$ for all $\boldsymbol{y_i}$ values, the $\boldsymbol{e_i}$ residual for each person will just be the difference between $\boldsymbol{y_i}$ and $\boldsymbol{\beta_0}$: $\boldsymbol{e_i} = \boldsymbol{y_i} - \boldsymbol{\beta_0}$
  - Rather than focusing on each individual $\boldsymbol{e_i}$ residual, we keep track of their **variance across persons** as the estimated model parameter, **denoted as $\boldsymbol{\sigma_e^2}$**
  - You've seen this before: $Variance = \boldsymbol{s^2} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1} = \frac{\sum_{i=1}^{N}(e_i)^2}{N-1} = \text{now } \boldsymbol{\sigma_e^2}$
    - In other words, the two parameters in the empty model give us the $\boldsymbol{y_i}$ outcome mean (as $\boldsymbol{\beta_0}$) and the y_i variance (as $\boldsymbol{\sigma_e^2}$) → right now $\boldsymbol{\sigma_e^2}$ = **all the $\boldsymbol{y_i}$ variance**

- In describing predictive linear models, the **notation refers to population parameters** instead of sample statistics (i.e., we use $\boldsymbol{\sigma_e^2}$ instead of s^2)
  - Why? Because we only ever have one sample from which to estimate parameters that we are trying to make inferences about with respect to some population

Beyond Empty GLMs: 2 Fixed Effects

- Purpose of predictive linear models (general and general*ized*) is to **customize** each person's expected outcome by adding **predictors**
  - Soon we will examine the unique effects of multiple predictors, but let's start with just one quantitative predictor: "(**simple**) **linear regression**"

- e.g., two quantitative variables, x_i and y_i, that both have a mean $(M) = 0$, a standard deviation $(SD) = 1$, and have a **Pearson's** $\boldsymbol{r = .5}$

- A GLM to describe **how** $\boldsymbol{x_i}$ **predicts** $\boldsymbol{y_i}$: $\boldsymbol{y_i = \beta_0 + \beta_1(x_i) + e_i}$
  - $\boldsymbol{\beta_1}$ = **slope** of x_i = difference in y_i per one-unit difference in x_i
    - $\boldsymbol{\beta_1 = r\left(\frac{SD_y}{SD_x}\right) = 0.5\left(\frac{1}{1}\right) = 0.5}$
  - $\boldsymbol{\beta_0}$ = **intercept** = expected y_i when $x_i = 0$
    - $\boldsymbol{\beta_0 = M_y - (\beta_1 * M_x) = 0 - (0.5 * 0) = 0}$

$\boldsymbol{\beta_1}$ is a linear slope (just like $\boldsymbol{r}$)

$\boldsymbol{\beta_0}$ adjusts for any mean difference between $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$

Unstandardized Intercepts and Slopes

- e.g., x_i and y_i both have $M = 0$, $SD = 1$, and $r = .5$
  - $\boldsymbol{\beta_1}$ = **slope** of x_i = still the difference in y_i per one-unit difference in x_i
    - $\boldsymbol{\beta_1 = r\left(\frac{SD_y}{SD_x}\right) = 0.5\left(\frac{1}{1}\right) = 0.5}$
  - $\boldsymbol{\beta_0}$ = expected y_i when $x_i = 0$
    - $\boldsymbol{\beta_0 = M_y - (\beta_1 * M_x) = 0 - (0.5 * 0) = 0}$

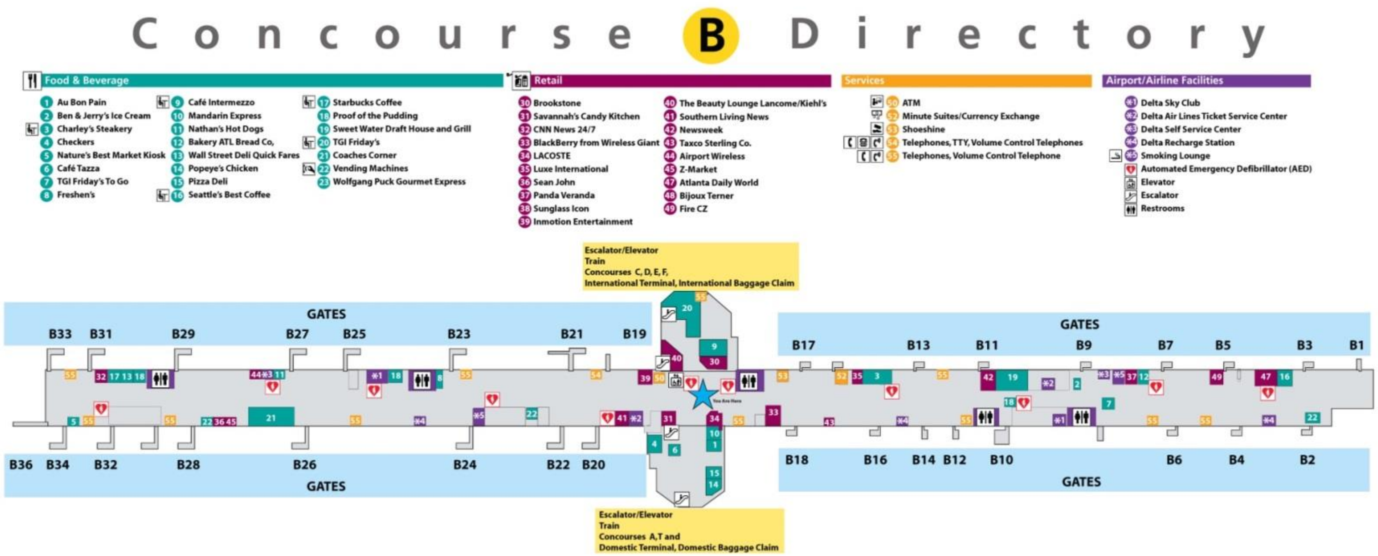
$\boldsymbol{\beta_1}$ is a linear slope (just like $\boldsymbol{r}$)

$\boldsymbol{\beta_0}$ adjusts for any mean difference between x_i and y_i

- What if $\boldsymbol{x_i}$ has $M = 50$, $SD = 10$ instead (but y_i still has $M = 0$, $SD = 1$)?
  - $\boldsymbol{\beta_1 = r\left(\frac{SD_y}{SD_x}\right) = 0.5\left(\frac{1}{10}\right) = 0.05}$
  - $\boldsymbol{\beta_0 = M_y - (\beta_1 * M_x) = 0 - (0.05 * 50) = 2.5}$

- What if $\boldsymbol{y_i}$ has $M = 50$, $SD = 10$ instead (but x_i still has $M = 0$, $SD = 1$)?
  - $\boldsymbol{\beta_1 = r\left(\frac{SD_y}{SD_x}\right) = 0.5\left(\frac{10}{1}\right) = 5.0}$
  - $\boldsymbol{\beta_0 = M_y - (\beta_1 * M_x) = 50 - (5.0 * 0) = 50}$

Why the Unstandardized Fixed Intercept β_0 *Should* Be Meaningful…



This is a very detailed map…
But what do we need to know
to be able to use the map at all?

Intercept ="You are Here" Sign

Using **original** years of education:

x_i =education, y_i = income

$$\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}(\boldsymbol{x_i}) + \boldsymbol{e_i}$$

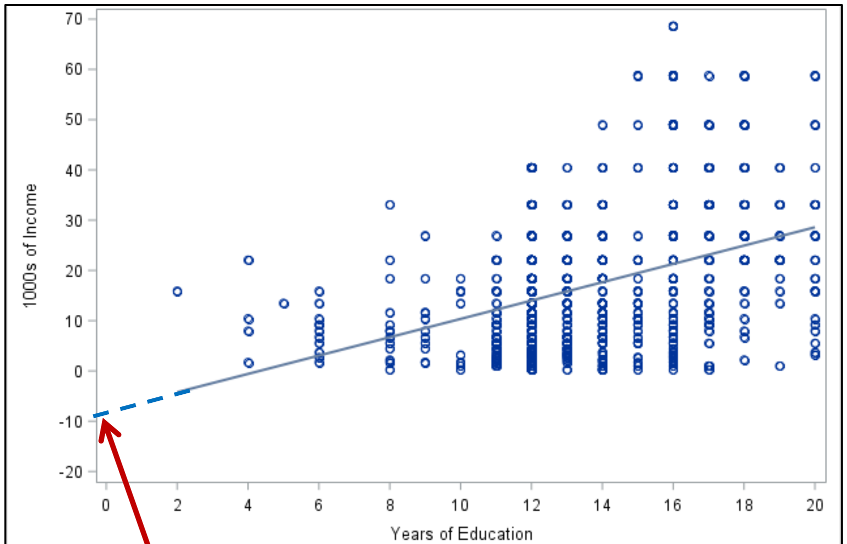
$$\boldsymbol{\hat{y}_i} = \boldsymbol{-7.89} + \boldsymbol{1.82}(\boldsymbol{x_i})$$

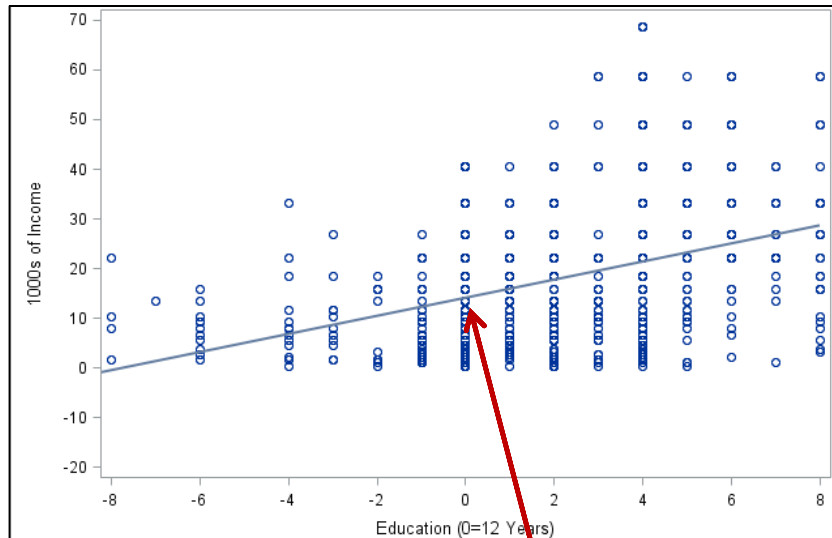
Using **education centered at 12**:

x_i =educ−12, y_i = income

$$\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}(\boldsymbol{x_i}) + \boldsymbol{e_i}$$

$$\boldsymbol{\hat{y}_i} = \boldsymbol{14.00} + \boldsymbol{1.82}(\boldsymbol{x_i})$$





Intercept $\boldsymbol{\beta_0}$

There is no *wrong* way to center, only *weird*. **Center so x_i=0 is meaningful.**

Intercept $\boldsymbol{\beta_0}$

Beyond Empty GLMs: Residual Variance

- Our GLM describes how x_i predicts y_i: $\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}(\boldsymbol{x_i}) + \boldsymbol{e_i}$
  - Intercept: $\boldsymbol{\beta_0}$; Slope of x_i: $\boldsymbol{\beta_1}$

- The $\boldsymbol{y_i}$ **expected** from the predictors is called $\boldsymbol{\hat{y}_i}$ = "**y hat**"
  - $\boldsymbol{\hat{y}_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}(\boldsymbol{x_i})$ → $\boldsymbol{y_i} = \boldsymbol{\hat{y}_i} + \boldsymbol{e_i}$ → $\boldsymbol{e_i} = \boldsymbol{y_i} - \boldsymbol{\hat{y}_i}$
  - Now we can determine what the $\boldsymbol{e_i}$ residual would be for each person, and thus what the variance of the $\boldsymbol{e_i}$ residuals would be

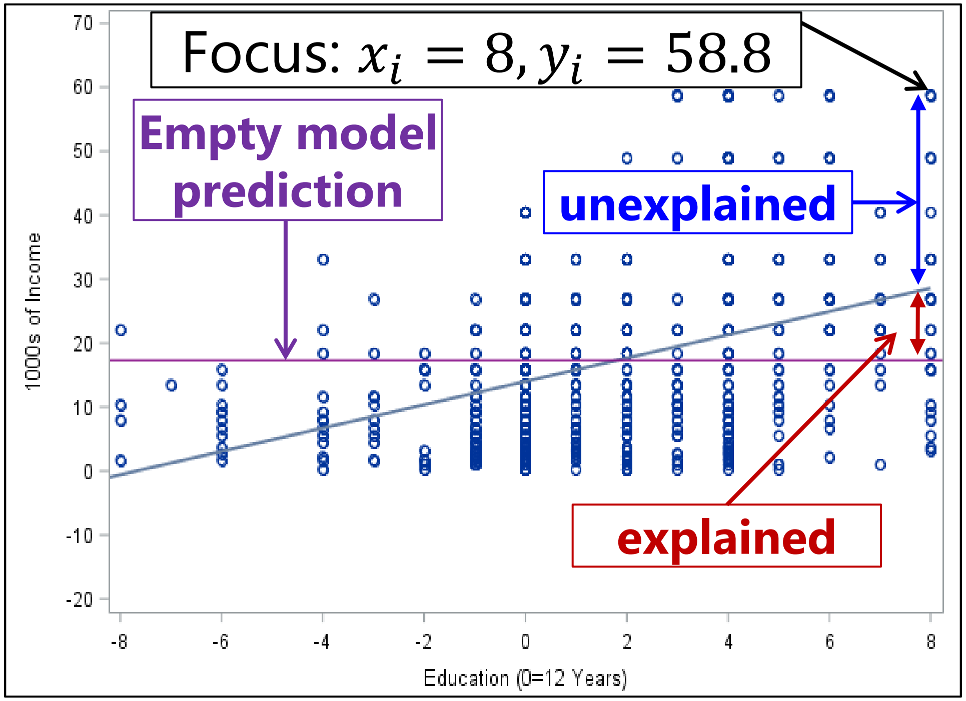
"**residual variance**": $$\boldsymbol{\sigma_e^2} = \frac{\sum_{i=1}^{N}(y_i - \boldsymbol{\hat{y}_i})^2}{N-2} = \frac{\sum_{i=1}^{N}(\boldsymbol{e_i})^2}{N-2}$$

  - Remember testing r against H_0 using the t-distribution with $N-2$? Same $N-2$ here, because we had to estimate two fixed effects: $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$

$$t = r\sqrt{\frac{N-2}{1-r^2}}\,,\ DF_{denominator} = N-2$$

More on GLM Residuals

The $\boldsymbol{\beta}$ formulas result from the goal of minimizing the squared residuals across the sample—this is called "**ordinary least squares estimation**"—let's see what happens for one example person



Empty Model for y_i = income:

$$\boldsymbol{y_i = \beta_0 + e_i}$$

$$\boldsymbol{\hat{y}_{Focus} = 17.3}$$

$$\boldsymbol{y_{Focus} = 17.3 + 41.5}$$

Variance: $\boldsymbol{\sigma_e^2} = \frac{\sum_{i=1}^{N}(y_i - \boldsymbol{\hat{y}_i})^2}{N-1} = 190.2$

→ 190.2 is **all** the y_i variance

Add Education as Predictor:

$$\boldsymbol{y_i = \beta_0 + \beta_1(Educ_i - 12) + e_i}$$

$$\boldsymbol{\hat{y}_{Focus} = 14.0 + 1.8(8) = 28.4}$$

$$\boldsymbol{y_{Focus} = 28.4 + 30.4}$$

Variance: $\boldsymbol{\sigma_e^2} = \frac{\sum_{i=1}^{N}(y_i - \boldsymbol{\hat{y}_i})^2}{N-2} = 162.3$

→ 162.3 is **leftover** y_i variance

Significance Tests of Fixed Slopes

- Each $\boldsymbol{\beta}$ **fixed slope** has 6 relevant characteristics to be reported:
  - **Estimate** = best guess for the fixed slope from our data
  - **Standard Error** = $\boldsymbol{SE}$ = average distance of sample slope from population slope → expected inconsistency of slope across samples
  - $\boldsymbol{t}$**-value** = (Estimate − H_0) / SE = test-statistic for fixed slope against $H_0(= 0)$
  - **Denominator DF** = $N - k$ (where k = total number of fixed effects)
  - $\boldsymbol{p}$**-value** = (two-tailed) probability of fixed slope estimate *as or more extreme* if H_0 is true → how unexpected our result is on a t-distribution with M=H_0, SD=SE
  - **(95%) Confidence Interval** = $\boldsymbol{CI} = Estimate \pm t_{critical} * SE$ = range in which true (population) value of estimate is expected to fall across 95% of samples

- Compare $\boldsymbol{t}$ test-statistic to t critical-value at pre-chosen level of significance (where % unexpected = alpha level): this is a "**univariate Wald test**"
  - Btw, if denominator DF are not used, then $\boldsymbol{t}$ is treated as a $\boldsymbol{z}$ instead
  - Because $\boldsymbol{\beta}$ **fixed slopes are unbounded**, SEs and CIs can be obtained directly (instead of through a Fisher r-to-z transformation as for r)

Significance Tests of Fixed Slopes

- **Standard Error** (**SE**) for the fixed slope estimate $\boldsymbol{\beta_x}$ in a single-predictor GLM: $\mathrm{SE}_{\beta_x} = \sqrt{\dfrac{\text{residual variance of Y}}{\text{variance of } x_i * (N-k)}} = \sqrt{\dfrac{\boldsymbol{\sigma_e^2}}{\sigma_x^2 * (N-k)}}$

- Example: $\boldsymbol{y_i = \beta_0 + \beta_1(Educ_i - 12) + e_i}$, $\boldsymbol{\sigma_e^2 = 162.28}$,
  $N = 734, x_i = Educ_i - 12: M = 1.81, Var = 8.46$

  - Slope for education: H_0: $\boldsymbol{\beta_1} = 0$, $\mathrm{H_A}$: $\boldsymbol{\beta_1} \neq 0$
    $Est = \mathbf{1.82}$, $\mathrm{SE} = \sqrt{\dfrac{\mathbf{162.28}}{\mathbf{8.46}*(734-2)}} = 0.16$, $t = \dfrac{Est-0}{SE} = \dfrac{1.82-0}{0.16} = 11.28$,
    $DF_{denominator} = N - k = 734 - 2 = 732$, $p < .0001$,
    $95\%\ CI = Est \pm (t_{crit} * SE) = 1.82 \pm (1.96 * 0.16) = 1.51$ to 2.14

  - Interpretation: Predicted income is **significantly higher** by 1.82k for each additional year of education (so reject H_0 that $\boldsymbol{\beta_1} = 0$)

SEs and CIs for Predicted Outcomes

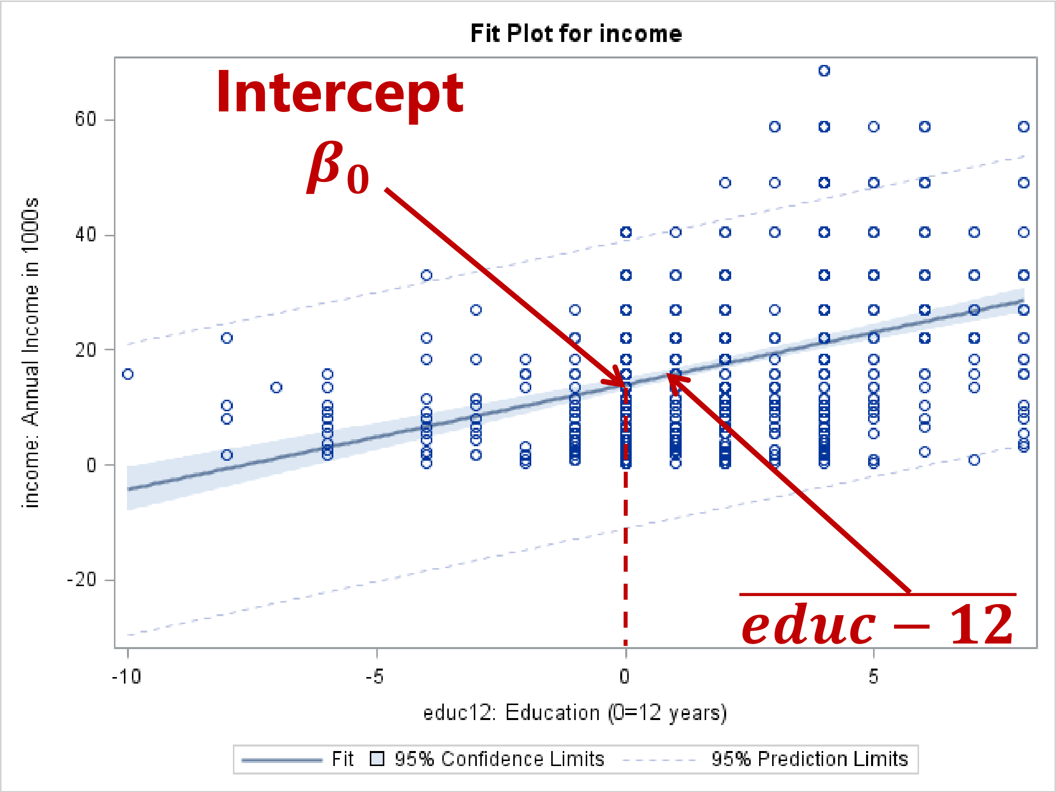
- The imprecision (SE) of any predicted outcome $\hat{y}_i$ (including the outcome captured by β_0) depends on the value of the predictor—the SE will increase as you move away from the predictor's mean:
  - SE of $\hat{y}_i \mid x_i = \sqrt{\sigma_e^2} * \sqrt{\frac{1}{N} + \frac{(x_i - \bar{x})^2}{(N-1)\sigma_X^2}}$

SE (for β_0 or any $\hat{y}_i$) = average distance of sample predicted value from population value

- $y_i = \beta_0 + \beta_1(Educ_i - 12) + e_i$, $\sigma_e^2 = 162.28$,
  $N = 734, Educ_i - 12: M = 1.81, Var = 8.46$
- SE and CI for predicted income when Education = 12?
  - Given by β_0: $Est = 14.00$, $\text{SE} = \sqrt{162.28} * \sqrt{\frac{1}{734} + \frac{(0-1.81)^2}{(733)8.46}} = 0.55$,
    $95\%\ CI = Est \pm (t_{crit} * SE) = 14.00 \pm (1.96 * 0.55) = 12.91 \text{ to } 15.09$

- You can use ESTIMATE in SAS or LINCOM in STATA to get predicted outcomes for any value of the model predictors…
  - Also options within each to get predicted outcomes for each person in data

CIs for Predicted Outcomes



Blue shaded line is created by $t_{critical} * SE$; blue dotted line also adds in error from σ_e^2

- The blue shading shows the 95% range for the $\hat{y}_i$ outcomes predicted by the regression line
  - They are narrowest at the predictor mean, and widen as moving away
- The blue dashed lines show the 95% range for the actual y_i outcomes implied by the residual variance (is way bigger)

Effect Size via Standardized Slopes

- GLM predictive equation uses the scale of the variables as entered directly into the model—this is the "**unstandardized**" solution
- e.g., $\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}(\boldsymbol{Educ_i} - \mathbf{12}) + \boldsymbol{e_i}$
  - x_i is $Educ_i - 12$: $M = 1.81, Var = 8.46$
  - y_i is $Income$: $M = 17.30, Var = 190.21$
- **Unstandardized:** $\boldsymbol{y_i} = \mathbf{14.00} + \mathbf{1.82}(\boldsymbol{Educ_i} - \mathbf{12}) + \boldsymbol{e_i}$
  - Unstandardized fixed slopes ($\boldsymbol{\beta_{unstd}}$) can be standardized ($\boldsymbol{\beta_{std}}$) as:

$$\beta_{std} = \boldsymbol{\beta_{unstd}} * \frac{SD_x}{SD_y}$$

$std\ \beta_0$ will always be 0

- **Standardized:** $\boldsymbol{y_i} = \mathbf{0} + \mathbf{0.38}(\boldsymbol{Educ_i}) + \boldsymbol{e_i}$
  - Standardized solution refers variables that have been transformed into $M = 0, Var = 1$ (i.e., as if they had been converted to z-scores)
  - Slopes are then in a familiar **correlation metric** (*usually* from -1 to 1)
  - Why do this? Standardized solution makes it **easier to compare the relative strength** of the fixed effects of predictors on different scales

What about Categorical Predictors?

- So far we've seen how a Pearson's r between two quantitative variables x_i and y_i can be represented equivalently with a general linear model of x_i predicting y_i: $\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}(\boldsymbol{x_i}) + \boldsymbol{e_i}$
  - Fixed slope $\boldsymbol{\beta_1}$ captures a linear effect of x_i predicting y_i in an unstandardized metric (using $\boldsymbol{x_i}$ centered so intercept at 0 makes sense)
  - For how to capture *nonlinear* quantitative predictor effects, stay tuned

- Now we will see how to use GLMs to predict a quantitative outcome from a **categorical predictor**
  - General rule: **predictors with $\boldsymbol{C}$ categories need $\boldsymbol{C}$ fixed effects** to distinguish the outcome means across all unique categories
    - After including the intercept $\boldsymbol{\beta_0}$, we still need $C - 1$ predictors, whose $\boldsymbol{\beta_x}$ slopes then capture mean differences between categories
  - So let's start with a **binary variable**, which requires a single predictor

A GLM with a Binary Predictor

- GLM of **binary** $\boldsymbol{x_i}$ predicting quantitative y_i:

$$\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}(\boldsymbol{x_i}) + \boldsymbol{e_i}$$

  - Create x_i so 0 = reference category, 1 = alternative category
  - Btw, this is called an "**Independent** (or **two-sample**) $\boldsymbol{t}$**-test**" (even though all types of predictors use a t test-statistic to test significance)

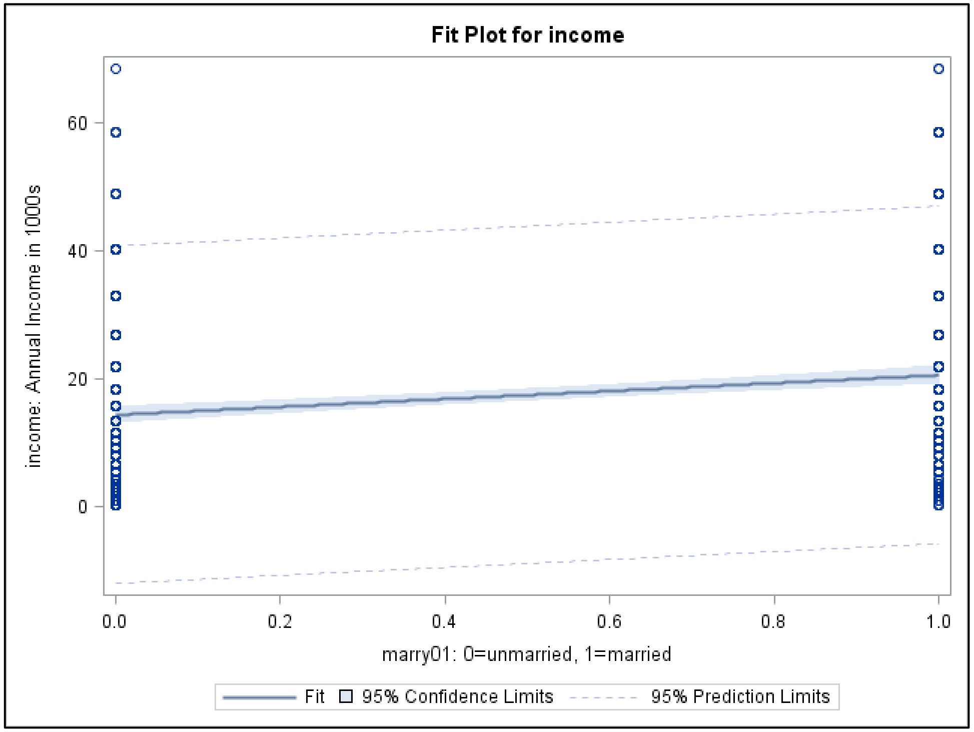
- For example: Family income predicted by marital status
  - $marrygroup_i$: 0 = no, 1 = yes → $\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}(\boldsymbol{Marry01_i}) + \boldsymbol{e_i}$
  - $\boldsymbol{\beta_0}$ = **intercept** = expected income for unmarried persons ($Marry01_i = 0$)
  - $\boldsymbol{\beta_1}$ = **slope** for $Marry01_i$ = expected mean difference for married persons relative to unmarried persons
  - $\boldsymbol{e_i}$ = **residual** = difference in model-predicted income (from $\boldsymbol{\hat{y}_i}$) and actual income $\boldsymbol{y_i}$, whose (residual) variance is estimated as $\boldsymbol{\sigma_e^2}$

A GLM with a Binary Predictor

$y_i = \beta_0 + \beta_1(Marry01_i) + e_i$ Income–marry $r = .23, p < .0001$

- Income predicted for unmarried:
  $\hat{y}_i = 14.45 + 6.22(0) = 14.45$
- Income residual for unmarried:
  $e_i = y_i - \hat{y}_i \rightarrow e_i = y_i - 14.45$

- Predicted income for married:
  $\hat{y}_i = 14.45 + 6.22(1) = 20.67$
- Income residual for unmarried:
  $e_i = y_i - \hat{y}_i \rightarrow e_i = y_i - 20.67$

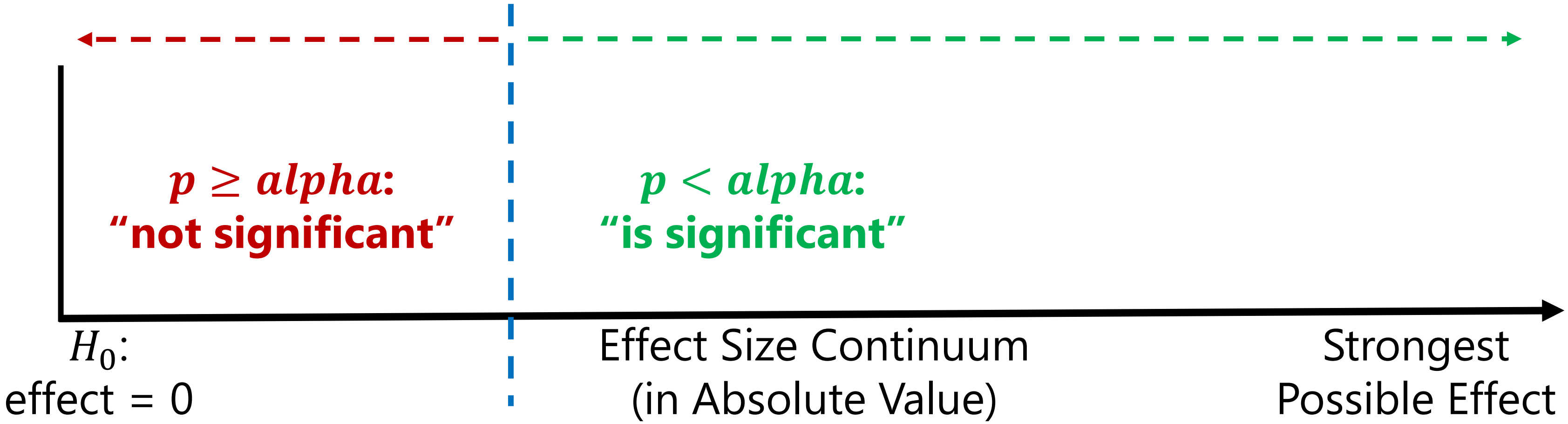


- A "linear" relationship is the only kind possible for binary predictors (there is only one possible "unit difference" in a binary x_i from 0 to 1)

Effect Size for a Mean Difference: d

- For categorical predictors, an $\boldsymbol{r}$ effect size (standardized slope) is less intuitive than an alternative effect size: **Cohen's $\boldsymbol{d}$**, a **standardized mean difference** between two groups (0 and 1)
  - $d = \frac{\bar{y}_0 - \bar{y}_1}{SD_{pooled}}$, where $SD_{pooled} = \sqrt{\frac{SD_0^2 + SD_1^2}{2}}$
  - Other variants you might see: Glass' delta (δ) uses SD for only 1 group; Hedges' g weights by the relative N in each group
  - If your GLM contains only one binary predictor, then the pooled SD is the same as the square root of residual variance, $\sqrt{\boldsymbol{\sigma_e^2}}$
  - Otherwise, $\sqrt{\boldsymbol{\sigma_e^2}}$ will be smaller because of the other predictors
    - d can be computed from t test-statistic for a fixed effect: $d = \frac{2t}{\sqrt{DF_{den}}}$
    - Btw, d and r can be converted as: $d = \sqrt{\frac{4r^2}{1-r^2}}$, $r = \sqrt{\frac{d^2}{4+d^2}}$

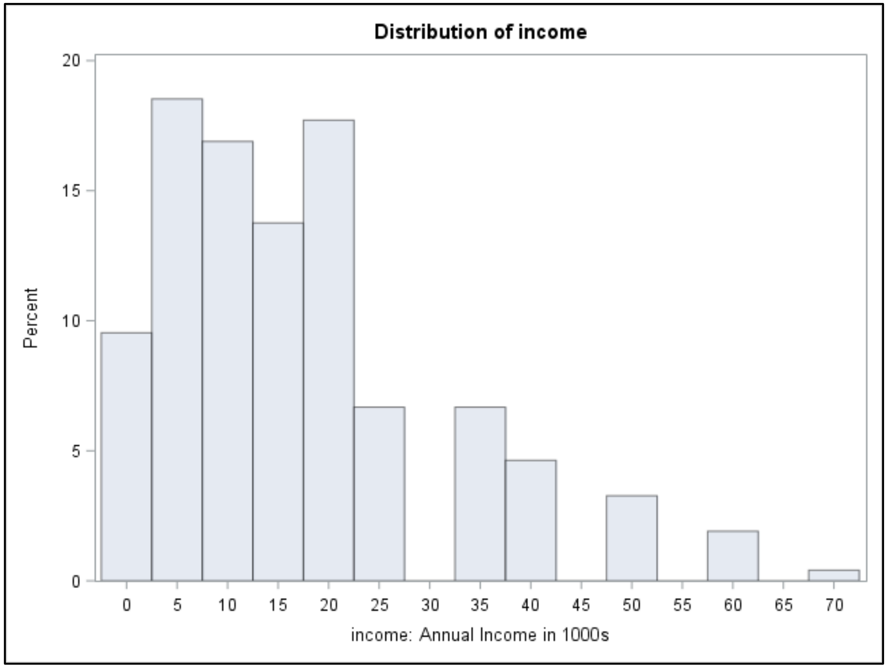
Effect Size, Sample Size, and Test Statistics

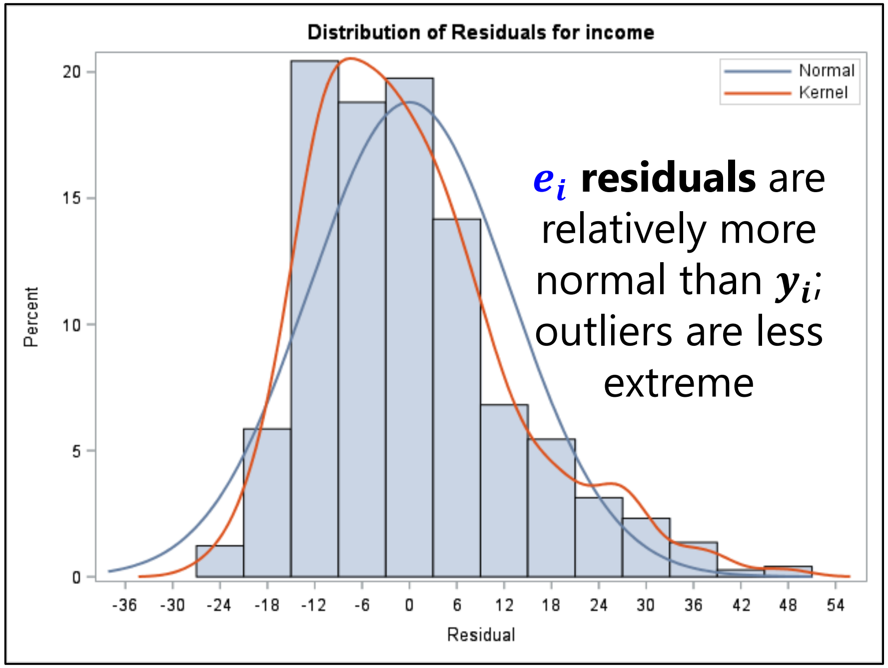


- Role of test statistics (t and F when using denominator DF; z and χ^2 if not) is to standardize a parameter's deviation from the null hypothesis
  - When compared to reference distribution, they give you a p-value: probability of finding an effect ≥ obtained effect **if H_0 is true**
  - **Test statistics** are a function of both **effect size** and **sample size N**
- In other words, test statistics and alpha combine to locate the blue line above that divides effect sizes into "not significant" and "significant"
- Blue line moves to the right (is harder to "find" an effect) given:
  - Lower alpha level = more conservative Type I error rate setting
  - Smaller sample size N → Fewer people = less power (higher Type II error)

What Choosing the GLM Means

- The GLM uses a **normal** distribution to describe the model outcome **residuals**, not the model *outcomes*—an important distinction!
  - That is, the **GLM specifies "conditional normality"** (of y_i given x_i)
- Our example: $\boldsymbol{y_i = \beta_0 + \beta_1(Educ_i - 12) + e_i}$
  - $\boldsymbol{\hat{y}_i = \beta_0 + \beta_1(Educ_i - 12)}$, so $\boldsymbol{y_i \sim N(\hat{y}_i, \sigma_e^2)}$ ← $\boldsymbol{y_i}$ is normally distributed with $M = \boldsymbol{\hat{y}_i}$ and $Var = \boldsymbol{\sigma_e^2}$





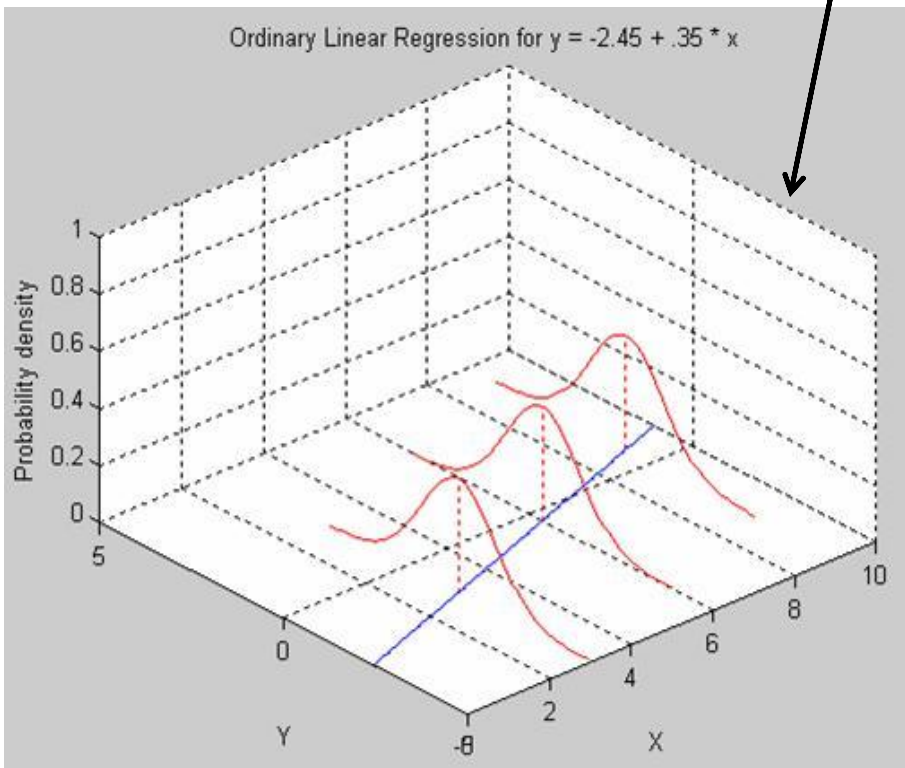
What Choosing the GLM also Means

- If **conditional normality** is not reasonable for your outcome, you may need to transform the outcome (meh, do so if you absolutely must) or choose a general*ized* linear model instead, otherwise your results (SEs and *p*-values) may be incorrect to some extent
  - Many outcomes cannot be transformed to become "more normal"
  - Come back in Spring 2022 for my **generalized linear models** class! (for categorical, binomial, count, and skewed continuous outcomes)

- Univariate GLMs also specify **independent e_i residuals**—that all the reasons why any pair of y_i outcomes would be more related than others are already accounted for in the model
  - Correlated ("dependent") residuals can result from sampling over more than one dimension (e.g., students from multiple schools)
  - Ignoring correlated residuals can lead to way-wrong results!
  - **Dependent residuals** require a "**multilevel**" or "**mixed-effects**" version of the general or generalized linear model instead (my other classes)

What Choosing the GLM also Means

- GLMs also specify equal (constant) residual variability across all predictor values: "**homoscedasticity**" = "**homogeneity of variance**"

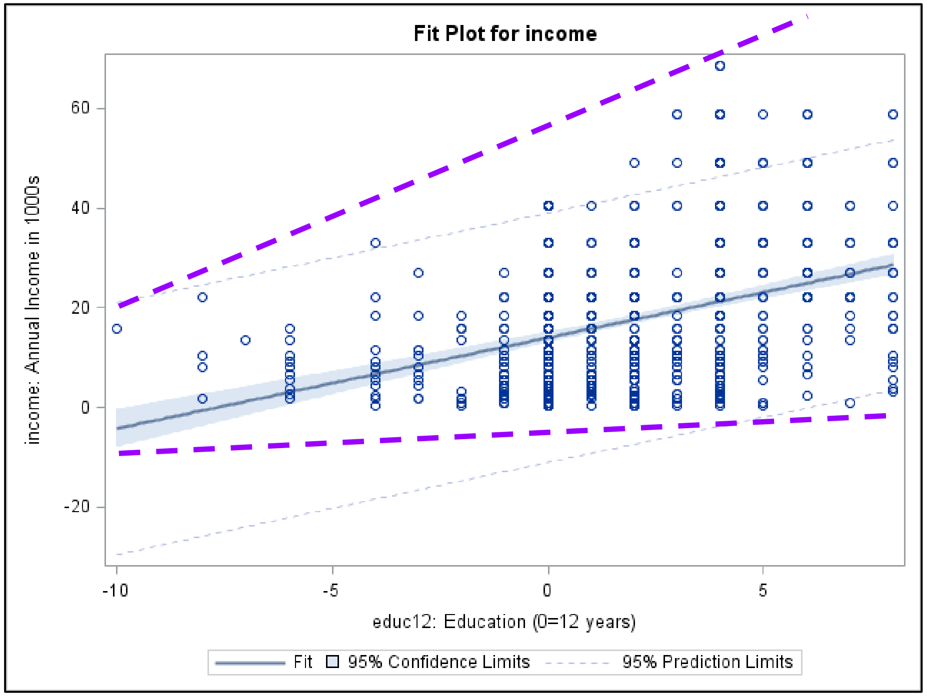


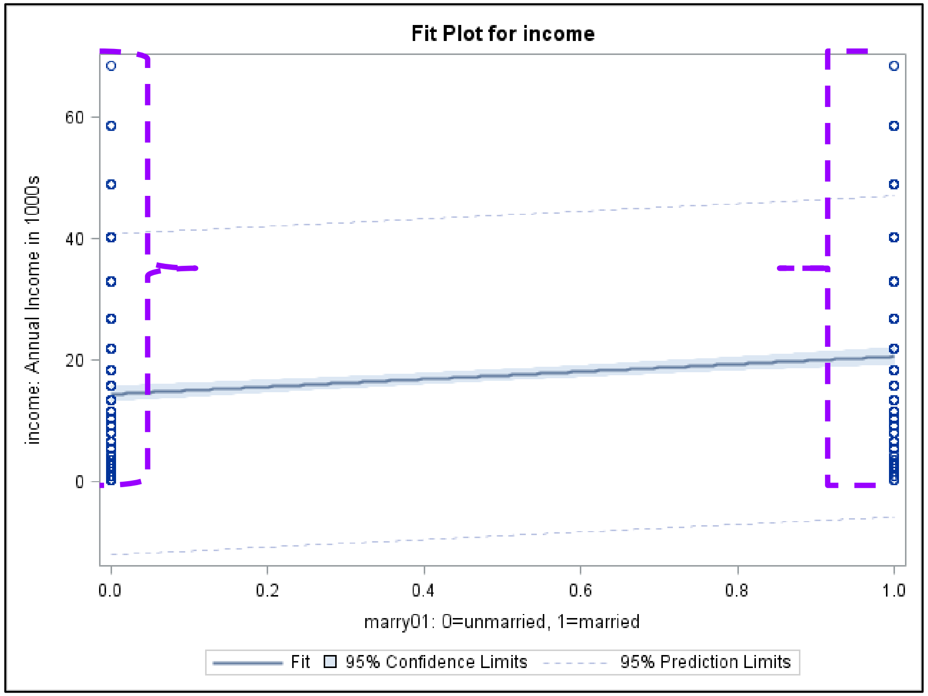
Otherwise, "**heteroscedasticity** " = "**heterogeneity of variance**" → model predicts differentially well across x_i (SE will need adjusted)

"Not good" → $\boldsymbol{\sigma_e^2}$ increases as the x_i predictor increases (→ fan shape)

Solution: Add fixed effects that allow the variance to differ (this leaves GLM)

Heterogeneity of Variance in Example Data





Left: Suspected heterogeneity of variance: Residual variance increases with education−12

Right: Apparent homogeneity of variance: Residual variance appears equivalent within married categories

Summary: Introduction to GLMs

- Predictive linear models (i.e., form as outcome = constant*predictor + constant*predictor…) create expected outcomes from 1+ predictors
  - **General** linear models use a **normal** conditional distribution
  - **General*ized*** linear models use **some other** conditional distribution

- General linear models are often called different names based on the type of predictor, but any kind of predictive model can be specified, for example:
  - **Empty Model**: no predictors; is used to recreate outcome mean and variance as unconditional starting point (sample mean is predicted for all)
    - $\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{e_i}$ → $\boldsymbol{\beta_0}$ = mean, $\boldsymbol{e_i}$ residual variance = $\boldsymbol{\sigma_e^2}$ → all the variance to be explained
  - **Single Predictor Model**: used to customize expected outcomes using a single predictor → $\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}(\boldsymbol{x_i} - \boldsymbol{C}) + \boldsymbol{e_i}$ (C is centering constant)
    - $\boldsymbol{\beta_0}$ = **intercept** = expected y_i when $x_i = 0$
    - $\boldsymbol{\beta_1}$ = **slope** of x_i = difference in y_i per one-unit difference in x_i
    - $\boldsymbol{e_i}$ = **residual** = deviation between actual y_i and predicted y_i (= $\boldsymbol{\hat{y}_i}$)
    - Effect size given by **standardized slope** will be equal to Pearson's r

- GLMs all specify residuals as normally distributed, independent, and with constant variance across predictors—otherwise, you need a new model!

Foreshadowing… please stay tuned!

- In a GLM with a **single predictor** (quantitative or binary), the effect size given by its **standardized slope** will be **equal to Pearson's r**
- So what's the point of estimating a GLM??? The real utility lies in **expanding the model** for at least one of these 3 reasons:
  - Multiple fixed slopes for a single predictor variable (in lecture 4)
    - To examine **nominal** or **ordinal predictors** of a quantitative outcome
    - To examine **nonlinear effects of a quantitative predictor** on a quantitative outcome (e.g., quadratic or piecewise spline predictors)
  - Multiple predictors (each potentially using 1+ fixed slopes)
    - To test the **unique effects** of correlated predictors after controlling for what information they have in common (coming in lecture 5)
  - Moderation of predictor effects (via interaction terms)
    - To test if predictor **slopes depend on** other predictors (lectures 6-7)