Stat 414 – Day 2
Adjusted Associations, Interactions, Modeling Heterogeneity

Last Time: Review of least squares regression

- Statistical model: $E(Y_i) = \beta_0 + \beta_1 x_i$, $\varepsilon_i \sim N(0, \sigma^2)$, $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$, $i = 1, \ldots, n$
  - Matrix form: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ with $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$
  - Least squares $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$
- Model assumptions:
  - Linearity between Y and *x*'s
  - Independence in the errors
  - Normality of the errors/Conditional response distributions at each *x* value
  - Equal variance of the errors/Conditional response distributions at each *x* value
  - X values are "fixed"
  - Outliers (large residuals)/Leverage (unique x-combo)/Influential observations
- Residual plots
  - Histogram/Probability plot of residuals
  - Residuals vs. Predicted values
    - No pattern (curvature, fanning, correlation with order number)
  - Residuals vs. individual explanatory variables, new variables
- Remedies: transformations, quadratic terms, new variables, additive models*
  - Centering can help with interpretation, reducing multicollinearity in related variables
- R^2 measures the proportion of variability in the response explained by the model
- R^2 adjusted includes a "penalty" for the number of explanatory variables in the model

Adjusted vs. Unadjusted Associations

Example 1: Data were collected on 654 youths in the area of East Boston during the middle to late 1970s. The youth in the study were of ages 3 to 19 years, an age period during which much physical development, such as increase in lung capacity, takes place. The objective was to analyze the relationship between smoking status, and forced expiratory volume (FEV, measured in liters). (FEV is a measure of strength of a person's lungs – the maximum volume of air a person can blow out in the first second; higher numbers are better.) The data can be found in the file **FEV.txt**.

(a) Run a regression model to predict FEV from smoking status. Write out the regression equation (using good statistical notation). How is the categorical variable being used in the model? How many df?

$\hat{FEV} = 2.566 + 0.71\ \text{smoker}_{yes}$

1 = smoker
0 = nonsmoker - reference group

(b) Interpret the slope coefficient in context.

On avg, [redacted] FEV is .71 larger for smokers than [redacted] nonsmoker

(c) Is the smoking variable statistically significant? How are you deciding? Df?
Note: This analysis is equivalent to a pooled *t*-test or a one-way ANOVA.

$t = 6.46$ $p = .0000000002$
reject $H_0: \beta_{smoker_{yes}} = 0$
equivalent $H_0: \mu_{smoker} = \mu_{nonsmoker}$
$F = 41.8 = (6.46^2)$
not causation because not randomized experiment

(d) Graph the residuals from your model in (a) vs age. Does there appear to be an association? In other words, might the age of the individual help explain some of the unexplained variation in FEV?

yes!

(e) Now fit the regression of FEV using both smoking status and age. Write out the full regression equation. How has the coefficient of the smoking status variable changed after adding age to the model? Interpret this coefficient in context.

$\hat{FEV} = 37 + .23\,age - .209\,Smoker\,yes$

.209 predicted decrease in FEV for smokers compared to nonsmokers of the same age

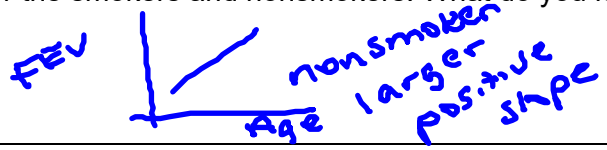
(f) Explain *why* the coefficient of smoking status changes. [*Hint*: What is the nature of the association between smoking status and age in this data set. Cite supporting output.]

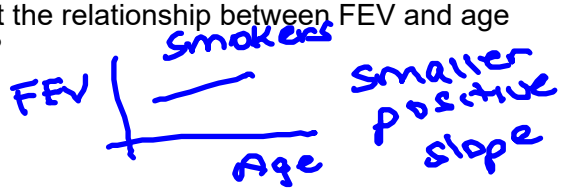
before really comparing older smokers (older ⇒ higher FEV) to younger nonsmokers

Interaction

Including the binary variable allows the intercepts to differ, but we are still assuming the slopes are the same.

(g) Produce a graph to decide whether there is evidence that the relationship between FEV and age differs for the smokers and nonsmokers. What do you learn?





Definition: A quantitative variable and a categorical variable ***interact*** if the slopes of the regression lines differ. (After all, it's the slope that tells us about the association between the two variables, so this says the association between the response and the quantitative variable depends on the category of the categorical variable.)

(h) Run the model including the interaction term (note the possible short-cut). Write out the full model equation. Is the interaction statistically significant? (State hypotheses, test statistic, df, p-value.) Is multicollinearity an issue?

$\hat{FEV} = .25 + 1.95\,smokeryes + .24\,age - .16\,smoker*age$

Smoker no: $\hat{FEV} = .25 + .24\,age$

Smoker yes: $\hat{FEV} = (.25 + 1.95) + (.24 - .16)\,age$

$H_0: \beta_{agesmoker} = 0$

reject $t = -5.29$, $df = 650$

(i) Center the age variable and rerun the model. Have things changed?

t & p-values for non interaction terms

(j) Summarize the nature of the interaction as if to a non-statistician. What does the interaction tell you about the "impact" of smoking on the relationship between age and FEV and what that means in this context?

To turn in

Homogeneity assumption Equal Variance

Example 2: Smith et al. (2005) examined reproductive and somatic tissues in the squid *Loligo forbesi*. The data in **Squid.txt** include the dorsal mantel length (in mm) and testis weight from 768 male squid, over different months. "The idea behind the original analysis was to investigate the role of endogenous and exogenous factors affecting sexual maturation, more specifically to determine the extent to which maturation is size-related and seasonal" (Zuur et al., 2009).

(a) Fit the model below, treating *Month* as a categorical variable

$$Testiswt_i = \beta_0 + \beta_1\, DML_i + \beta_2\, Month_i + \beta_3\, DML_i \times Month_i + \varepsilon_i$$

Examine the residual plots. What do you conclude?

unequal variance (increasing w/ $\hat{y}$)

We can consider transformations (but unlikely to be helpful here?). We can also consider modelling the variance. A common choice is the variance is related to an explanatory variable.

(b) Is there evidence that the variance is related to either explanatory variable here?

The variance in the response appears to increase with our explanatory variable DML. So we can instead model $\varepsilon_i \sim N(0, \sigma^2 \times DML_i)$

We can think of this as a special case of ***weighted least squares*** which assumes $V(Y_i \mid X_i) = \sigma^2/w_i$ and if we know the w_i then we minimize $\Sigma w_i (Y_i - \beta_0 - \beta_1 X_i)^2$.

(c) First, let's use $1/w_i = DML_i$. What should be the impact of using these weights on the least squares estimates? (What does the covariance matrix look like?)

$$\begin{bmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{bmatrix} \rightarrow \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{bmatrix}$$

Less weight for squid w/ large DML → pay more attention to small squid

(d) Run the weighted least squares estimation (model 2). What has changed? (Also consider residual standard error and R^2?)

$\hat{\sigma} = 2.55 \rightarrow .1505$

(e) Have things improved? Let's check the residuals.

not really

Note: To examine residual plots for assessing the model assumptions with weighted regression, you would need to use the "weighted residuals" $\sqrt{w_i}\,(y_i - \hat{y}_i)$. These weighted residuals will be reflected if you look at the *standardized* (aka *normalized*) residuals, which divide the residuals by by "s", instead of the raw residuals. looked a little better

Now, let's forget about the increasing variance with *DML* but allow the variances to vary by month. We can write this model as follows:

$$\text{Testisweight}_{ij} = \text{intercept} + DML_{ij} + \text{Month}_j + \text{DML}_{ij}{:}\text{Month}_j + \text{residuals}_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma_j^2) \quad j = 1, \ldots, 12$$

(f) First, we want to create a new variable that we can use as the weights. We will assign the variance of the data from each month to that month.

σ_3^2

$s_3^2 = 3.075$

(In fact, each variance estimate is a multiple of the baseline group's estimate.)

(g) Run the weighted least squares regression using these weights. Does it help?

yes (look at standardized residuals)

But there is a different problem with how we have done things here. We don't know σ_j, we have only estimated it from the sample data. Instead we should use ***generalized least squares*** (Aiken, 1934) which is an iterative approach for simulataneously estimating the fixed regression coefficients and estimating the variance terms.

(h) Use R's *gls* command to create model 3b (models 1b and 2b are identical to above).

(i) Fit model4b which allows the power of DML to vary by month.

Notes:

- Can formally test the equality of the variance estimates, controlling for the "fixed effects" (but need to learn more about GLS and maximum likelihood estimation first)
- Can explore other "forms" (e.g., powers) of the *variance covariates*.
  - Can be issues with zero and negative values
- Can allow relationship with variance covariate to differ across categories varIdent(form=~x|cat)
- Can include interactions of categorical variables to allow for different variances at each factor-level combination
- You can use the fitted values rather than one of the explanatory variables form=~fitted(.)