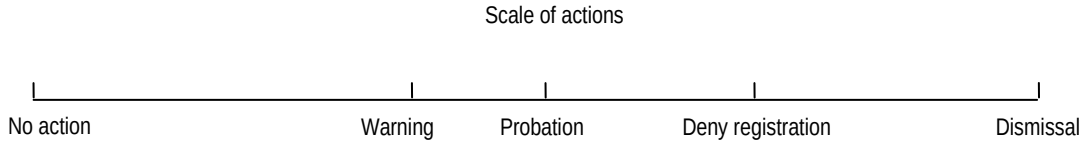
Section 6 Functional Form and Nonlinearities

This is a good place to remind ourselves of Assumption #0: That all observations follow the same model.

Levels of measurement and kinds of variables

- There are (at least) three essential kinds of variables in econometrics
    - **Interval (or cardinal) variables** are the "usual" kind of variables that are continuous and where the numbers actually measure something.
        - Differences are meaningful: An income of $70,000 exceeds an income of $60,000 by the same amount as an income of $50,000 exceeds an income of $40,000
        - With interval variables, we can talk meaningfully about continuous mathematical functions and partial derivatives
    - **Ordinal variables** take on several ordered values, but differences are not meaningful.
        - It is a scale on which we know which direction different values are from one another, but not how far each adjacent pair is apart.
        - Example: highest academic action taken against a student.
            - We know that dismissal is worse than denial of registration, denial is worse than probation, probation is worse than warning, and warning is worse than no action.
            - We don't know *how much* worse each is than the adjacent action. Denial may be a bigger step from probation than probation is from warning.



        - With ordinal variables, we can talk meaningfully about how a change in another variable would move this variable along its scale, but there is no easy single-number translation into how that movement on the scale would translate into the underlying ordinal levels of the variable because we don't know a priori how far they are apart. (Econometrics allows us to estimate this if we use proper procedures.)
        - Are GPA and SAT scores interval or ordinal?
    - **Categorical variables** are those that have several possible outcomes but where those outcomes cannot even be ranked ordinally.

        - For these variables, we just have to treat the outcomes as separate possibilities and cannot meaningfully put them on a scale at all.
        - Example: Choosing to attend Reed vs. another school would be a two-outcome categorical variable. (Choosing Reed vs. L&C vs. another school would be three-outcome variable.)
        - Sex, ethnicity, and many other variables are categorical.

Dummy (binary or indicator) independent variables

- Dummy variables are (yes, no) variables. We traditionally give the value 1 to yes and 0 to no.
- Dummy variables are used to model categorical variables as dependent or explanatory variables and ordinal variables as explanatory variables.
    - When there are only two possible outcomes, a single dummy variable is sufficient (e.g., sex, ignoring the transgendered).
    - When there are $M > 2$ outcomes, we need $M - 1$ dummy variables:
        - Region $\in$ {Northeast, South, Midwest, West}
        - Need dummies for Northeast, South, Midwest.
        - Don't need dummy for West because we can tell those observations from the fact that they are zero for the other three.
        - If we include all four dummies, they will add up to 1, meaning perfect multicollinearity in a regression that also includes an intercept term.
- While dummy variables are often very useful in multiple regressions (with more than one regressor), they are limited in simple regression, but have a special interpretation.
    - Suppose that D is a dummy variable for sex with $D = 1$ being male.
    - Consider the model $y_i = \beta_1 + \beta_2 D_i + e_i$. For females, $D = 0$ and the expected value of y is β_1. For males, $D = 1$ and the expected value of y is $\beta_1 + \beta_2$. Thus, β_1 is the difference between the expected y for males and females.
    - A test of the null hypothesis $\beta_2 = 0$ would be a test of whether males and females have the same average y.
        - This is equivalent to the t test for the equality of means and is a simple application of "analysis of variance."
- When there are other variables present, a dummy variable shifts the intercept of the relationship between y and the other variables upward or downward depending on the value of the dummy. (Slope is assumed to be the same.)
- Consider the following example: $\ln W = \beta_1 + \beta_2 ED$ to estimate the effect of an additional year of education on the wage.
    - Wages may differ across sexes
        - To allow the function to vary by a constant amount between males and females: $\ln W = \beta_1 + \beta_2 ED + \beta_3 MALE$

        - For females: $\ln W = \beta_1 + \beta_2 ED$
        - For males: $\ln W = (\beta_1 + \beta_3) + \beta_2 ED$
        - Thus, the intercept is different for males than for females, but the slope is the same, meaning that we have assumed that the effect of education on wages is the same for males and females.
    - What if we include a *FEMALE* dummy as well?
        - $\ln W = \beta_1 + \beta_2 ED + \beta_3 MALE + \beta_4 FEMALE$
        - This would add nothing of value to the regression because we already know the difference between males and females from β_3.
        - $MALE + FEMALE = 1 = x_1$ so there is perfect multicollinearity: **X′X** is singular and the inverse does not exist.
        - Statistical algorithms will either break down or (like Stata) delete one of the collinear variables.
    - How would this work with the regional dummies?
        - To model differences in intercept, we include dummies for three of four regions: $\ln W = \beta_1 + \beta_2 ED + \beta_3 Northeast + \beta_4 South + \beta_5 Midwest$
            - Again, including all four results in collinearity
        - The intercept term β_1 is the intercept for the *omitted category* (West)
        - The intercept for South is $\beta_1 + \beta_4$, so β_4 measures whether the intercept is different for the South vs. the West.
        - Choose the omitted category to be the one against which you want to test others, then the *t* test is easier
        - To test whether region matters at all, do a joint *F* test of $\beta_3 = \beta_4 = \beta_5 = 0$.
- If we have a dummy variable that is 1 for only a single observation (presumably in multiple regression), then the residual for that observation will be zero and the coefficient of that dummy variable will have the value of the residual of that observation in an otherwise identical regression that excludes the dummy.
- Dummy dependent variables can be estimated by OLS using the **linear probability model**, but this is not the best way to estimate these models, so we won't go into any details.

Interaction effects

- What if the effect of education differs for males and females?
    - In this case, we need an **interaction variable**.
    - $\ln W = \beta_1 + \beta_2 ED + \beta_3 MALE + \beta_4 (ED \times MALE)$
    - Equation for females: $\ln W = \beta_1 + \beta_2 ED$
    - Equation for males: $\ln W = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) ED$

- Thus, β_3 is the difference in the intercept and β_4 is the difference in the slopes.
- Running this regression is equivalent to running separate regressions for the male and female samples
  - The female sample will have an intercept estimate of b_1 and a slope estimate of b_2
  - The male sample will have an intercept estimate of $b_1 + b_3$ and a slope estimate of $b_2 + b_4$
  - Running them together requires that the variance of the error term for males and females be the same
  - But running them together allows testing the hypotheses $\beta_3 = 0$ and $\beta_4 = 0$, which are often of interest.
  - The joint test that both (all) coefficients are the same across the two subsamples is called a **Chow test**.

o Allowing the slope (education effect) to vary across regions would involve interaction terms between *ED* and each of the three regional dummies.

o We can also interact dummies with one another:

$$\ln W = \beta_1 + \beta_2 ED + \beta_3 MALE + \beta_4 South + \beta_5 \left(MALE \times South \right)$$

- For non-South females: $\ln W = \beta_1 + \beta_2 ED$
- For South females: $\ln W = \beta_1 + \beta_4 + \beta_2 ED$
- For non-South males: $\ln W = \beta_1 + \beta_3 + \beta_2 ED$
- For South males: $\ln W = \beta_1 + \beta_3 + \beta_4 + \beta_5 + \beta_2 ED$
- South effect for females = β_4
- South effect for males = $\beta_4 + \beta_5$
- Male effect for non-South = β_3
- Male effect for South = $\beta_3 + \beta_5$
- Thus, β_5 measures the difference in the male effect between South and non-South, or the difference in the South effect between males and females.

- We can also **interact continuous variables**
  - $\ln W = \beta_1 + \beta_2 ED + \beta_3 AGE + \beta_4 \left(ED \times AGE \right)$
  - $\dfrac{\partial \ln W}{\partial ED} = \beta_2 + \beta_4 AGE$
  - $\dfrac{\partial \ln W}{\partial AGE} = \beta_3 + \beta_4 ED$
  - Thus, β_4 measure the effect of age on the value of an additional year of education, or the effect of education on the value of an additional year of age.

Treatment effects

- Correlation does not imply causation, even if one event occurred before the other.
    - Hospital stay vs. health status example from text
    - **Selection bias** occurs when the sample of people is not randomly chosen from the population.
        - Wage equation example: only working people have observed wages, but people with higher wage offers are more likely to work
- **Randomized controlled experiments** are gold standard of statistical procedures, but are not often available in economics.
    - Unless we have the resources to create our own data, we must use **natural experiments** arising out of natural variation in observed datasets.
    - Randomized experiments randomly place observations into **treatment group** or **control group**.
    - Often use "double-blind" technique in medical treatments where neither the patient nor the doctor knows which group the patient is in: avoiding the **Hawthorne effect**.
- Problems with experiments:
    - Lack of randomization can lead to correlation between group selection and other variables.
        - Can control for this by controlling for these variables by including them in a regression.
    - Partial compliance
        - Did the treatment and control groups actually do what they were supposed to do?
        - Did the job-training selectees actually attend training?
        - Did the patient take the drug?
        - Is this behavior correlated with e?
    - Attrition
        - Some drop out of both groups during the experiment.
        - Were they random or did people with high (or low) values of e drop out?
    - Hawthorne effect
        - Double-blind is not possible in many experiments.
        - Experimenter bias may result from incentives to make results look significant.
- **Difference estimator**
    - Let $d_i = 1$ for observations in treatment group, 0 for those in control group.
    - $y_i = \beta_1 + \beta_2 d_i + e_i$
    - β_2 measures the "treatment effect"

- OLS regression will give $b_2 = \bar{y}_1 - \bar{y}_0$, the differences of the means of the two groups.
- Do we need other regressors?
  - Not if selection is random because there is no omitted variable bias if the omitted regressors are uncorrelated with the variable of interest (*d*).
  - If selection is non-random, but *all* the variables that determine the selection are observable and added to the regression, then our dummy-variable coefficient will still be unbiased because there are no omitted variables that are correlated with *d*.
  - If we allow people to "select into" the treatment and control groups, then there will be other characteristics (those that affect the choice) that will be correlated with *d*. If any of these variables are also correlated with *y*, then we have omitted variable bias ~ sample selection bias.
  - It may still be useful to include other regressors because they will lower the overall variance of the equation and reduce the amount of variation that *d* needs to explain.
  - Interaction terms between treatment dummy and other regressors would allow us to see how treatment effect might vary with subject characteristics.

- **Differences-in-differences estimator**
  - When we only have natural experiments, we can sometimes do before and after comparisons between the control and treatment groups and get valid estimators under appropriate conditions.
  - In order to do this we must have two observations (before and after) for each unit and we must be able to assume that the before→after change is independent of any omitted variable correlated with treatment status.
  - The differences-in-differences estimator uses the model $y_{it} = \beta_1 + \beta_2 d_i + \beta_3 t + \delta(d \times t) + e_{it}$, where $t = 0$ for "before" and 1 for "after" and *d* is the treatment dummy we used above.
  - $$\hat{\delta} = \left(\bar{y}_{treatment,\,after} - \bar{y}_{control,\,after}\right) - \left(\bar{y}_{treatment,\,before} - \bar{y}_{control,\,before}\right)$$
    $$= \left(\bar{y}_{treatment,\,after} - \bar{y}_{treatment,\,before}\right) - \left(\bar{y}_{control,\,after} - \bar{y}_{control,\,before}\right)$$
  - You can also use other controls to reduce variance here
  - Differences-in-differences estimator is example of using **panel data**, which vary both across time and across units.
    - We will study methods for use with panel data latter on.

Nonlinearity in variables vs. nonlinearity in parameters

- Solving for the OLS estimator required that we differentiate the LS or likelihood function with respect to the parameters.
- In a model that is linear in parameters, the LS objective function will be quadratic, so that the least-squares normal equations based on setting the first derivatives to zero are linear in the coefficient estimator.
    - This means that we can use linear algebra to solve for the coefficient estimator.
- If the model is nonlinear in parameters, then the LS objective function will not be quadratic and the normal equations will not be linear in parameters, so numerical search methods must be used for solution.
    - This is called **nonlinear LS** and is much more computationally difficult and potentially problematic than the linear model. (Covered in S&W appendix to Ch. 8.)
    - There are times when nonlinear LS is necessary, but we try to avoid it whenever possible.
- There are many models that are nonlinear in variables but linear in parameters. These models are easy to deal with: we can transform the variables and use linear OLS methods.
- If a model is nonlinear in its regressors (or with a nonlinear dependent variable), then the coefficient on the variable is no longer $\partial y/\partial x_j$.
    - Instead, we have to calculate $\partial y/\partial x_j$ as a function of the coefficients and the values of X.
    - This will vary according to the functional form, so we'll talk about the partial effects for individual forms as we discuss them.
- The choice of functional form should be guided by theory, but theory rarely provides a unique specification.
    - It is often necessary to try various functional forms to see which one seems to fit the best.
    - Plotting actual and fitted values against each regressor can often be helpful in seeing nonlinearities.
- One way to explore nonlinearities (if you have a large enough sample) is to create a battery of dummy variables with different levels of a regressor. Looking at the pattern of coefficients for the different levels can tell you whether the relationship is approximately linear.
    - For example, we could examine math SAT score effects by looking at dummies for $500 \leq \text{SATM} < 600$, $600 \leq \text{SATM} < 700$, and $\text{SATM} \geq 701$, leaving out the bottom category below 500.
    - This will give us four points on a general response function (with zero implicit for the omitted group, below 500).

- If the four points seem to lie on a straight line, then the linear specification is probably fine. One may also see evidence of quadratic or cubic behavior and can use more than four categories if you have enough data and want to be more discriminating.

Quadratic and higher-order polynomial models

- One easy way of incorporating curvature into a model is to introduce quadratic terms. (For the moment, we will assume only one regressor is nonlinear, so we'll ignore others.)
  - $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$
  - Possible shapes for the relationship:
    - Upward sloping at an increasing rate ($\beta_1 > 0$, $\beta_2 > 0$)
    - Upward sloping at a decreasing rate or downward sloping but flattening out ($\beta_1 > 0$, $\beta_2 < 0$)
      - Note that this curve *always* turns downward (upward) after a peak (trough) at $x = -\beta_1/2\beta_2$, so it is critical to evaluate which part(s) of the curve the sample lies in. (Are most/all of the x values of interest < or > $-\beta_1/2\beta_2$?)
      - This non-monotonicity may be good or bad depending on theory.
      - If you want a universally monotonic but diminishing effect, using ln x may be a good alternative specification.
    - Downward sloping and getting steeper as x increases ($\beta_1 < 0$, $\beta_2 < 0$)
  - Always include a graph of the response function so that your reader can understand the shape of the effect.
    - The coefficients don't tell the story in a transparent way.
  - Partial effect
    - $\frac{\partial Y}{\partial X} = \beta_1 + 2\beta_2 X$.
    - The sign of the partial effect will change at $x = -\beta_1/2\beta_2$ if $\text{sgn}(\beta_1) \neq \text{sgn}(\beta_2)$, as discussed above.
  - Estimating the standard error of the partial effect
    - Conditional on x,

      $$\text{var}\left(\hat{\beta}_1 + 2\hat{\beta}_2 X\right) = \text{var}\left(\hat{\beta}_1\right) + 4X^2 \,\text{var}\left(\hat{\beta}_2\right) + 4X \,\text{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right).$$

      The estimated values of the variances and covariances can be obtained from the output of your regression package. (They are the diagonal and off-diagonal elements of the estimated covariance matrix of the coefficient vector. This is obtained by estat vce after a regression command. (As usual, it will be the classical estimated covariance estimator unless you use the robust option in the regression.)

        - S&W point out two other ways of estimating the standard error of a linear combination of coefficients:
            - Do a test command that the partial effect is zero to get an F statistic, then an estimate of the standard error will be the absolute value of the partial effect at that x divided by the square root of the F value.
            - Transform the model into one where the desired effect is directly estimated and get the standard error from the regression table.
    - Relevant significance tests in the quadratic model:
        - Does x affect y?
            - This is a test of the joint null hypothesis $H_0 : \beta_1 = 0, \beta_2 = 0.$ It is a standard F test.
        - Is the relationship quadratic rather than linear?
            - This is a t test of H_0: $\beta_2 = 0$, given that β_1 is assumed to be nonzero (null hypothesis is linear model).
            - This is an example of a **nested specification test** because the linear model is a special case of (nested within) the quadratic specification.
            - Note that the t test is preferred to comparing R^2 or $\bar{R}^2$ values.
                - The former will always be higher for the quadratic specification.
                - The latter will be higher if the t value exceeds one, which is well below conventional critical values.
- Higher-order polynomials
    - Do cubic, quartic, etc. relationships ever occur in economic data?
        - Yes, but they can be hard to sell.
        - Example of SAT scores and Reed GPA.
    - Same procedures apply for estimated partial effects and tests.
        - What to do if 3rd-order term is significant and 2nd-order term is not?
            - Don't leave out the 2nd-order term.
            - Test both jointly to try to reject the linear model in favor of the cubic. If significant, retain both.

Nonlinear least squares

- For models that are nonlinear in the parameters, we must generally use nonlinear search methods to find the least-squares (or maximum-likelihood) estimates.
    - Linearity in parameters depends crucially on the specification of the error term.
    - The error term in the model *must* be additive.
        - Consider the model $y = e^{\beta_1} x^{\beta_2}$.

- If the appropriate error term specification is $y = e^{\beta_1} x^{\beta_2} e^{e}$, then we can take logs and get $\ln y = \beta_1 + \beta_1 \ln x + e,$ which is linear in the parameters and can easily be estimated by linear OLS.
- If the appropriate error term specification is $y = e^{\beta_1} x^{\beta_2} + e$, then the model cannot be made linear in parameters with an additive error term and must be estimated nonlinearly. (I don't know why this error specification would be better, but just suppose…)

- Nonlinear estimation usually requires you to (at minimum) provide a formula for the deterministic part of the function.
  - To estimate the above model in Stata you could type
    nl (y = exp({b1}) * x^{b2}), initial (b1 5 b2 0)
  - Nonlinear search algorithms can be very slow and unreliable. It is generally very helpful to provide starting values near the optimal parameter values.
    - In this case, we might run the log-log regression (using the wrong error term specification) to get preliminary estimates of the coefficients, then insert those values in the "initial" option of the nl statement.
- Nonlinear estimation is a directed search over the parameter space to find the best combination. It is generally guided by taking numerical derivatives of the objective (LS or likelihood) function with respect to the parameters, then following the direction of greatest improvement (the gradient).
  - Some nonlinear-optimization packages allow you to enter analytic (algebraic) partial derivatives of the model with respect to the parameters. This generally speeds up convergence.
- Some objective functions may have multiple local optima. Starting far from the global optimum can cause the algorithm to become trapped at a global optimum that is inferior to the global one. Good initial values can help avoid this problem.
  - To assure that your optimum is a global one, try starting from several different sets of initial values and see if you converge to the same optimum.
- Some objective functions are badly behaved, having ridges (or valleys) where the objective function is very flat in one direction. This is particularly true if multicollinearity is a problem. If two variables are highly, positively correlated, then increasing the coefficient of one by a lot and simultaneously decreasing the coefficient of the other will have very little effect on the predicted values and the residuals, hence on the objective function. This leads to a ridge in the likelihood function (valley in the least-squares function) at a diagonal in the space of these two variables.
- Nonlinear estimation is not as computationally problematic as in the old days, but it is still subject to these numerical difficulties.
  - Avoid it when possible by using specifications that are linear in the parameters.

- There will be times when we need to use it for maximum-likelihood estimators such as probit and logit, but these likelihood functions are often well-behaved.