EPSE 596: Correlational Design and Analysis

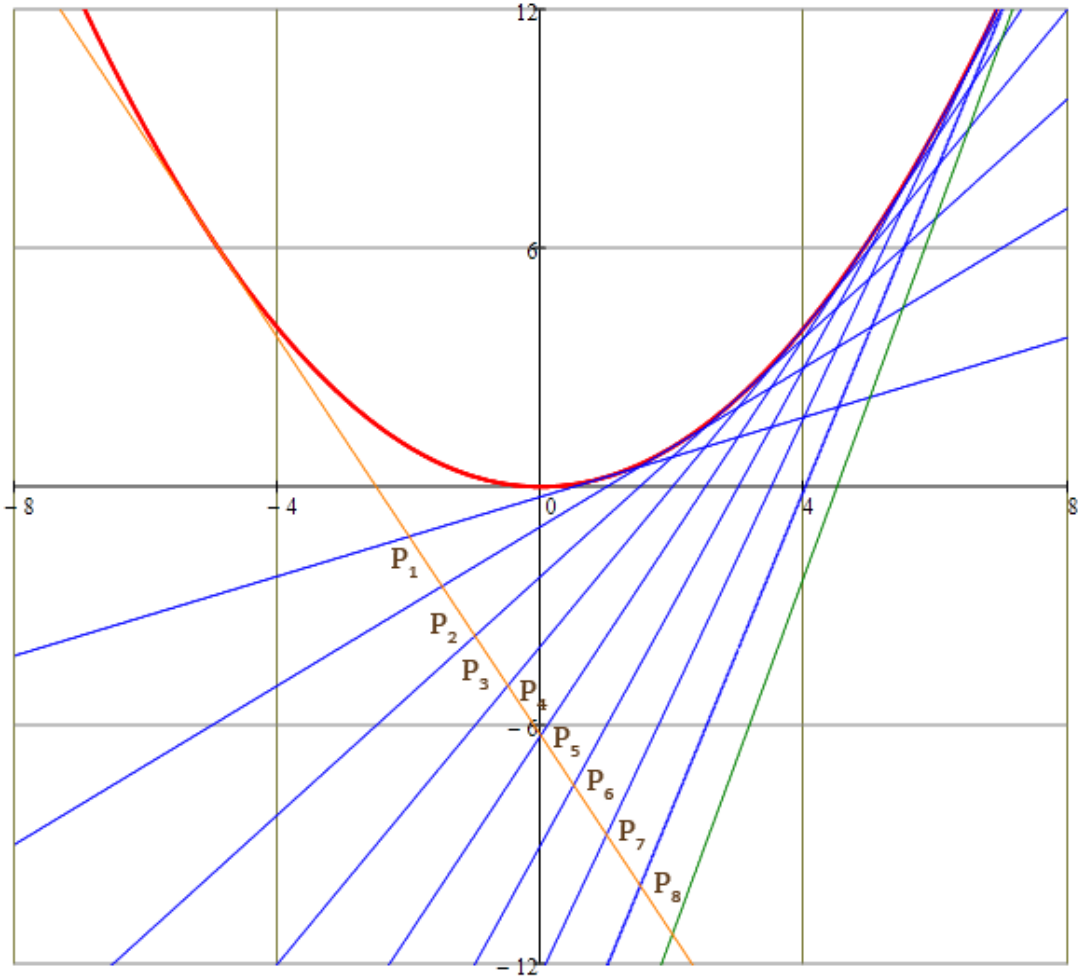
Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

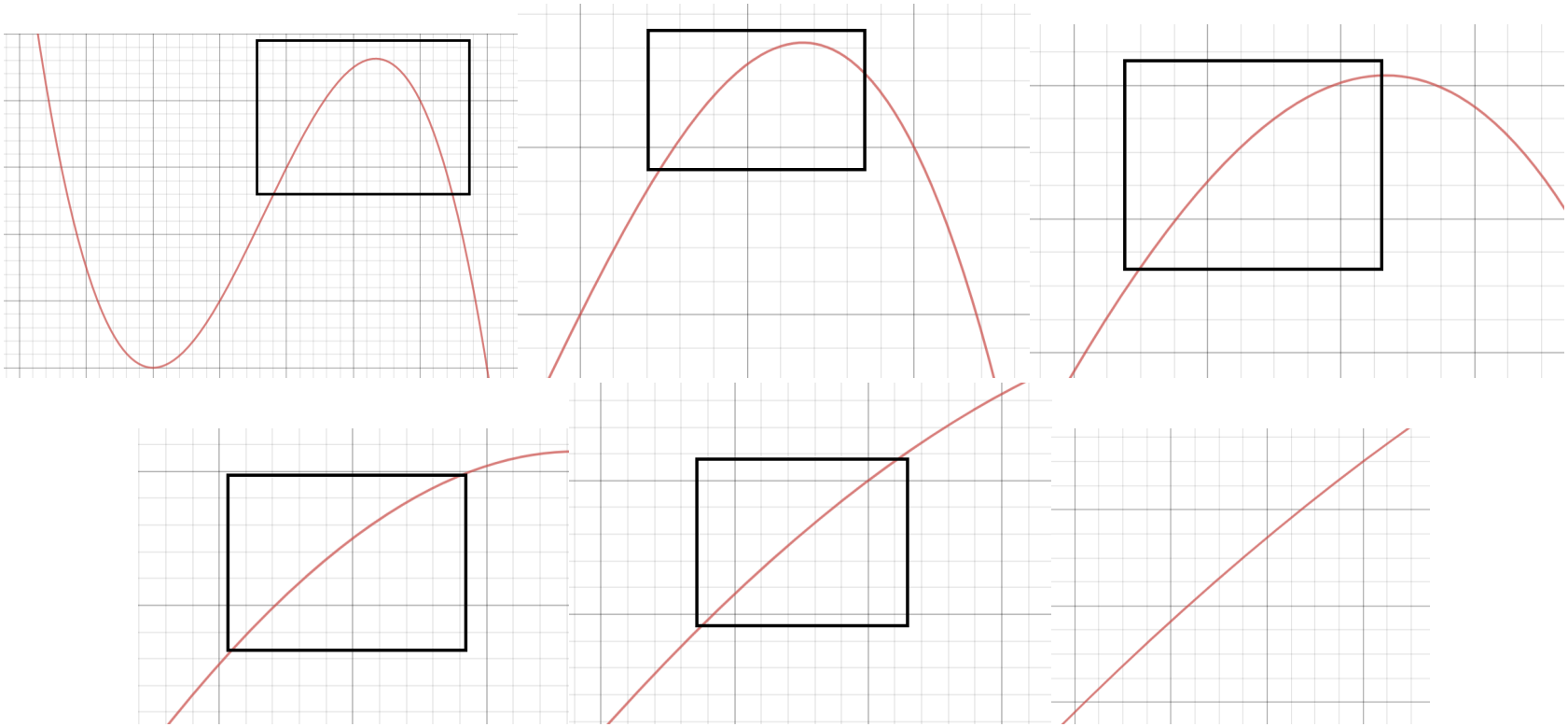
Incorporating curvature into a model

When an (X, Y) relationship is curvilinear, we can always approximate the relationship by a line at a *local* level. Here, $Y = X^2$, so the slope of the *local linear approximation* is $2X$.



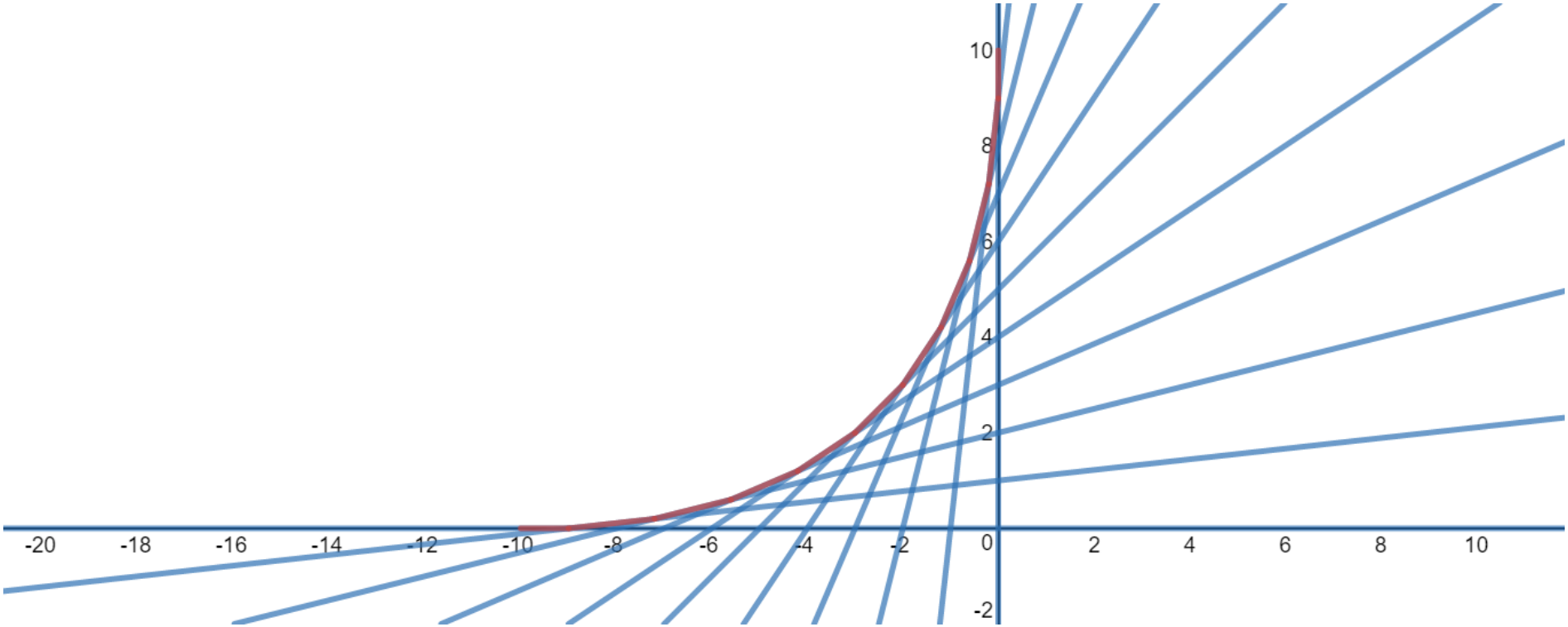
Incorporating curvature into a model

For most curves, if we zoom in enough, the (X, Y) relationship will always look linear; hence, always have *local linearity*.



Incorporating curvature into a model

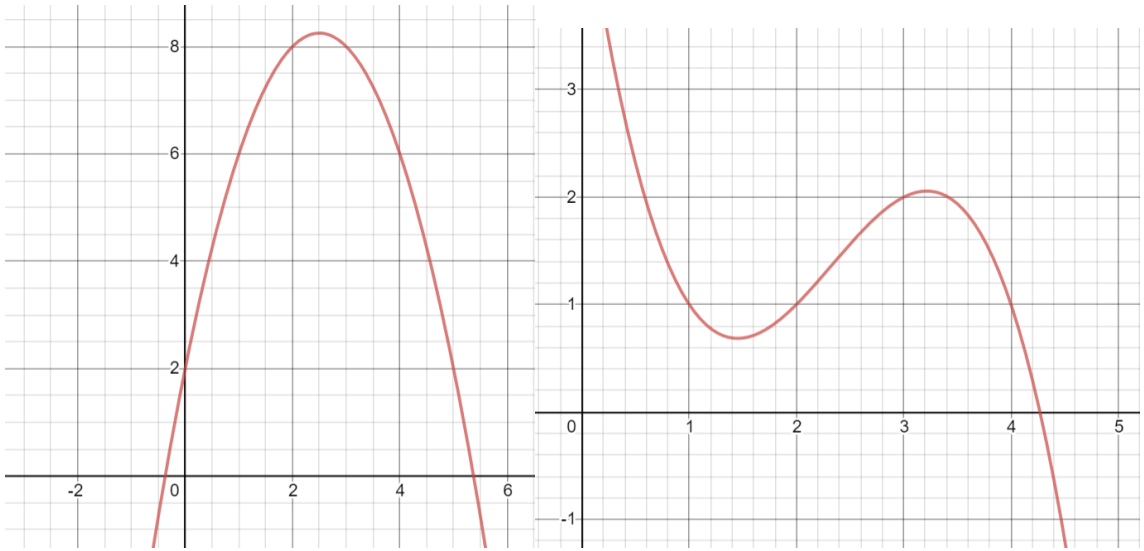
The slope of the *local linear approximation* depends on the particular algebraic relationship between (X, Y) globally. Here, $Y = 8e^{x/2}$, so the slope of the local linear approximation is $4e^{x/2}$.



Note: The slope of the local linear approximation is a consequence of basic calculus (i.e., it is the *derivative*).

Incorporating curvature into a model

- In practice, you can (almost) always get away with considering a *polynomial* relationship between X and Y in your regression model. This is because nonpolynomials are often well approximated by polynomials (Weierstrass Approximation Theorem).
- In fact, for most small and/or medium sized datasets, you can get away with considering only a *quadratic* or *cubic* relationship between X and Y.



Quadratics and cubics are enough

- In practice, with real data, quadratic (i.e., second order) models like

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2)$, are great when the (X, Y) relationship *levels out at one boundary* of your X-predictor data:

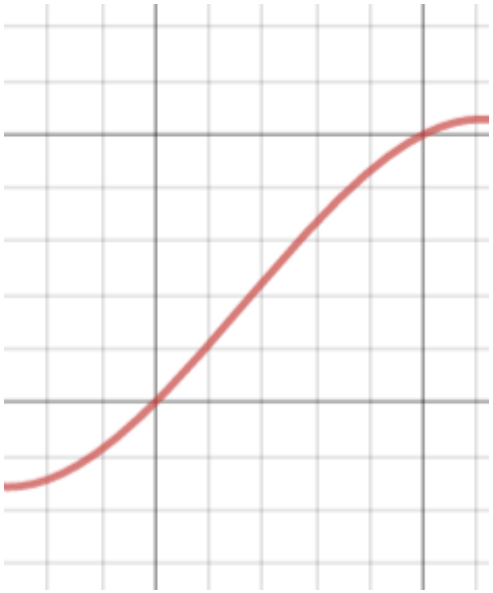
Slope of local linear approximation is: $\beta_1 + 2\beta_2 X$

Quadratics and cubics are enough

- In practice, with real data, cubic (i.e., third order) models like

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2)$, are great when the (X, Y) relationship *levels out at both boundaries* of your X-predictor data:



Slope of local linear approximation is: $\beta_1 + 2\beta_2 X + 3\beta_3 X^2$

Quadratics and cubics are enough

- In practice, with real data, cubic (i.e., third order) models like

$$g(\mu_Y) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3,$$

are great when the $(X, g(\mu_Y))$ relationship *is roughly parabolic*:

Slope of local linear approximation is: $\beta_1 + 2\beta_2 X + 3\beta_3 X^2$

Incorporating curvature into a model

- The best things about incoporating curvature into your model:
  - Can capture finer details in the relationship between predictor(s) and response.
  - More realistic model (often).
  - Can always plug coefficients of regression model into the formula(s) for the local linear approximation and so interpretation of effects are the same as before (i.e., unit change in X leads to a $f(\beta_X, X)$ change in Y on average.
- Downsides about incorporating curvature into your model:
  - Must work harder to interpret effects at a local level (although, reality is complicated! This is probably necessary work).
  - Need more data to estimate more model coefficients.
  - Need more data to reliably estimate higher order effects.

Incorporating multiple predictors into a model

In most real world problems, there will be *more than one* predictor that could be incorporate into your model.

- Sometimes referred to as *multiple regression* when more than one distinct predictor variable.
- Model assumptions do not change, but now the relationships are *higher dimensional*.
- That is, every time you add another unique predictor into the model, you add *another dimension* to the phenomenon you are studying. For example,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$$

can be described in *two* dimensions, while

$$Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \varepsilon$$

requires *three* dimensions.

Incorporating multiple predictors into a model

The most typical kind of second-order model you will likely consider though is one that includes an *interaction* between X_1 and X_2:

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

- Unlike first-order models or the other second-order models with marginal curvature that we have seen, this interaction model captures a linear relationship between X_1 and Y that is *mediated/moderated/tempered* by the other predictor X_2. Also, the linear relationship between X_2 and Y is mediated/moderated/tempered by the other predictor X_1.
- That is, for a unit increase in X_1, we expect Y to change on average by $\beta_1 + \beta_3 X_2$. Notice that the *strength of the linear association is tempered by the value of* X_2.
- Similarly, for a unit inrease in X_2, we expect Y to change on average by $\beta_2 + \beta_3 X_1$.

Incorporating multiple predictors into a model

The most typical kind of second-order model you will likely consider though is one that includes an *interaction* between X_1 and X_2:

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

- We will return to the "mediation/moderation" language in a later week, since the social sciences (especially psychology) and some of the health sciences use this terminology to describe certain interaction effects between variables.
- I will try to *avoid* this terminology though; an *interaction* between two variables is simply when the correlation/effect between the response and one of the interacting variables changes depending on the value of the second interacting variable.

Incorporating multiple predictors into a model

We can continue to build more complex models by incorporating more unique predictors into our models, each time adding another dimension of complexity to what we are trying to model.

- Three predictors with two pairwise interactions:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \varepsilon$$

- Three predictors with marginal curvature in X_1, two pairwise interactions, and a three-way interaction:

$$Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_1^2 + \gamma_3 X_2 + \gamma_4 X_3 + \gamma_5 X_1 X_2 + \gamma_6 X_1 X_3 + \gamma_7 X_1 X_2 X_3 + \varepsilon$$

- Five predictors with no curvature or interactions:

$$Y = \varphi_0 + \varphi_1 X_1 + \varphi_2 X_2 + \varphi_3 X_3 + \varphi_4 X_4 + \varphi_5 X_5 + \varepsilon$$

Incorporating multiple predictors into a model

- The best things about incoporating additional unique predictors into your model:
  - More realistic models.
  - Can capture how different variables *interact* with each other to explain variation in a response phenomenon.
  - Can still interpet regression coeffiicents as before, although now we speak of average change in Y for a unit increase in X, *keeping all other unique predictors fixed*.
- Downsides about incorporating additional unique predictors into your model:
  - Must work harder to interpret effects (although, reality is complicated! This is probably necessary work).
  - Need more data to estimate more model coefficients.
  - Need more data to reliably estimate higher order effects (including interactions).

Interactions can mimic curvature

- Notice: If two predictors X and Z are at all *correlated* (usually so in practice), and if the interaction $X \cdot Z$ helps explain variation in your model, then it is possible that X^2 and/or Z^2 could help explain at least as much *or more* variation in your model instead.
- So you can have situations where you are *spuriously* modelling an interaction between two predictors that is better understood or explained as marginal curvature in one of the predictors.
- Especially in situations where you have moderate or large correlation between predictors (colinearity – will return to this in a future class), if you are finding an interaction between the predictors to improve validity or fit of the model, consider if the validity and/or fit would improve even more if you included a quadratic term in one of the predictors instead.

Putting it all together to build a model

- Let's work through an example with some real data to try to build a good model to explain variation in a response Y with three unique predictor variables: X, W, and Z.
- **There is no one, correct way to go about building a good model.** Lots of trial and error. Lots of exploration. Some good general things to do (or not do):
  (1) Plot the response vs. each predictor individually.
  (2) If any predictor variables should be in your model because of prior knowledge and/or theoretical considerations, then *put them in the model*.
  (3) Do *not* just remove predictors if the estimated coefficients are not significant (i.e., not statistically distinguishable from zero). Will wreak havoc on the quality of your model.
  (4) Always check fitted vs. residual plots for clues about model misspecification and validity.

Putting it all together to build a model

(5) Always examine *residual standard error*; hopefully, it should *decrease* as your model improves.

(6) Examine simpler models (e.g., marginal curvature in one variable at a time, or two predictor with an interaction model) to see if they explain some variation before building more complex models that incorporate these simpler models.

(7) Remember: Your estimated model coefficients will *change as you change the functional form of your model*. Hence, your estimated effects are **model-dependent**, *not just data-dependent*.

(8) For real research settings, be prepared to spend *many hours* considering different regression models.

(9) **NEVER** trust *automated model building* procedures unless you have huge datasets on relatively few predictor variables.