LINEAR REGRESSION

Outline

- About Regression
- Mechanics of Estimation
- Prediction and Inference
- Models

Linear Regression

- Linear regression is used to model the relationship between a numeric outcome variable and one or more features by fitting a linear equation to the data.
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{\mathrm{p}} x_{\mathrm{p}} + \varepsilon$$
- The model estimates coefficients that minimize the sum of squared residuals.
- Linear Regression is one of the oldest predictive modeling techniques.
- Many modern machine learning approaches are generalizations or extensions of linear regression

Use Cases

- Predicting housing prices from square footage and location
- Estimating insurance premiums from age and health indicators
- Forecasting company revenue based on marketing spend
- Estimating blood pressure based on age and BMI
- Forecasting electricity demand from temperature and time of day

Strengths

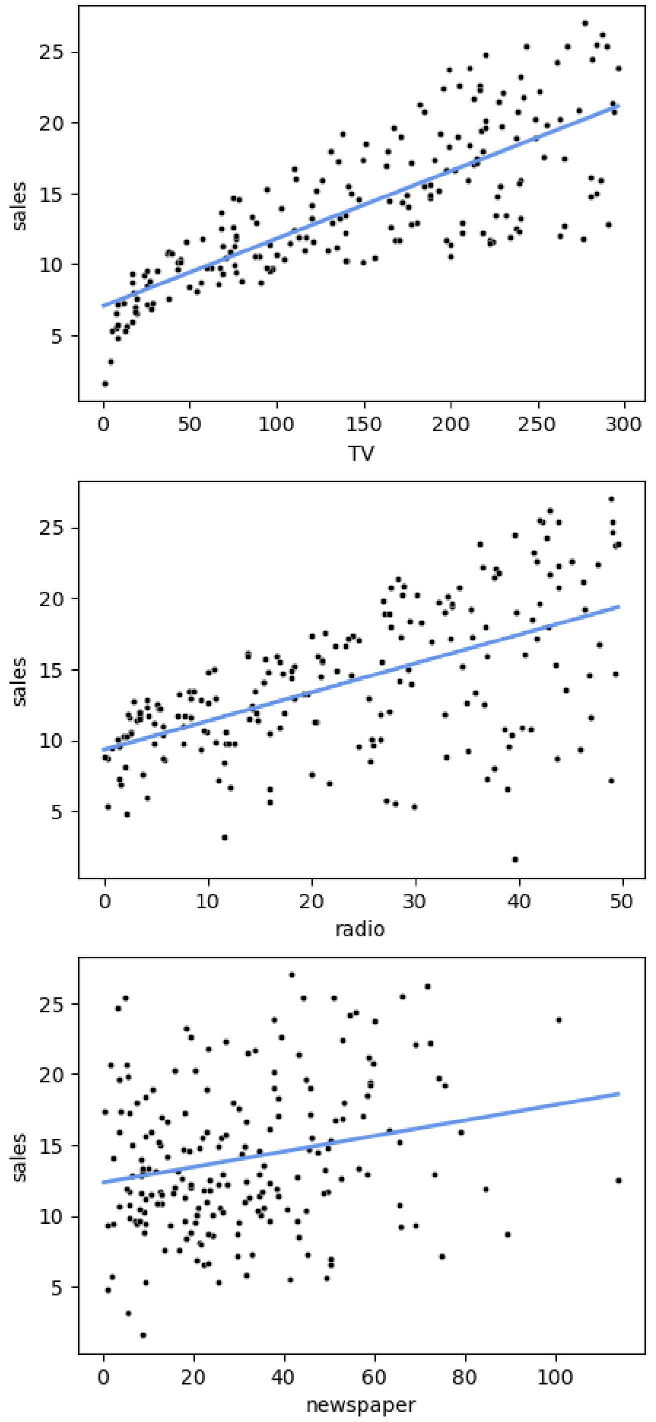
- Well-understood and extensively used
- Fast and computationally efficient
- Easy to interpret and communicate

Weaknesses

- Assumes a linear relationship between features and outcome
- Sensitive to multicollinearity, outliers, and missing data
- Cannot capture complex non-linear interactions unless transformed
- Makes a number of assumptions such as constant variance and normally distributed errors

Questions Regression May Answer Based on Advertising Data

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?



Represents similar figure from Introduction to Statistical Learning with Applications in Python, 2023

Regression

1. Estimate Regression Equation
2. Prediction
3. Inference

Let’s begin by examining the estimation process

MECHANICS OF ESTIMATION

Estimate Regression Equation

- Estimate parameters of the population regression equation
- $Y = \beta_0 + \beta_1 X + \varepsilon$
  - *where X is the predictor,*
  - *Y is the outcome,*
  - *β_0 and β_1 are regression coefficients*
  - *ε is random error*
- Coefficients estimated using an optimization procedure like Ordinary Least Squares (OLS)
  - *Construct a linear combination of predictors such that $\Sigma e_i = 0$ and Σe_i^2 is minimum*
- Next few slides will illustrate this optimization process using an example.

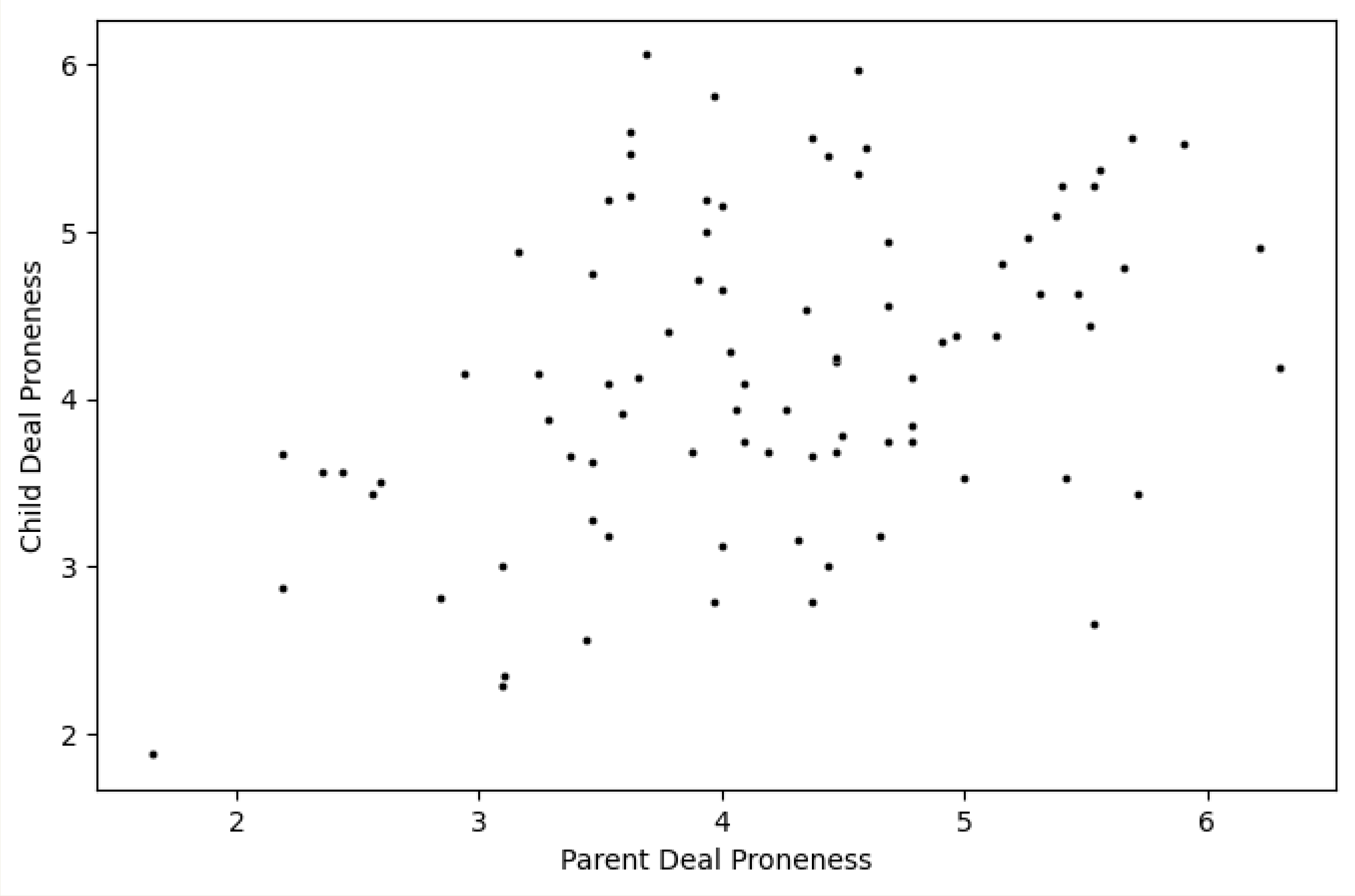
Example

- Deal proneness is the tendency of shoppers to buy products that that offer a good deal such as coupon discounts, sales and buy-one get-one free offers.
- Does deal proneness of parents affect deal proneness of children?
- Schindler, Lala, and Grussenmeyer (2014) gathered data on deal proneness of parents and their children using a 32-item scale for deal proneness. The scores were averaged to construct an index.

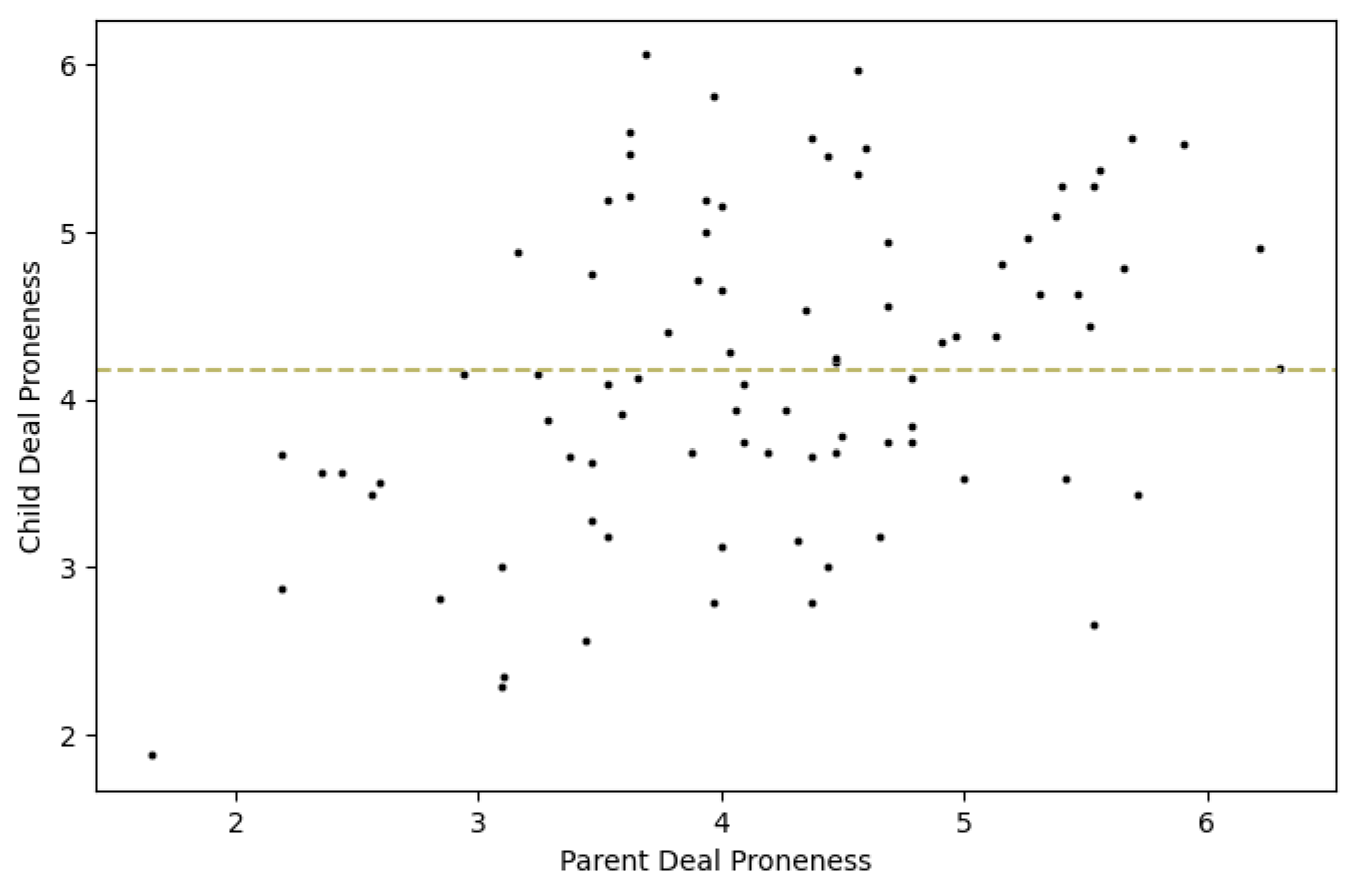
Source: Schindler, Robert. M., Vishal Lala, and Colleen Corcoran (2014). “Intergenerational Influence in Consumer Deal Proneness,” Psychology & Marketing, 31 (5), 307-320

id	Parent (X)	Child (Y)
1	5.0	3.5
2	3.9	5.0
3	5.5	4.6
4	3.4	2.6
5	3.6	5.6
6	5.9	5.5
7	2.6	3.5
8	5.7	3.4
9	4.4	2.8
10	4.1	3.9
..	..	..
..	..	..

Scatter Plot

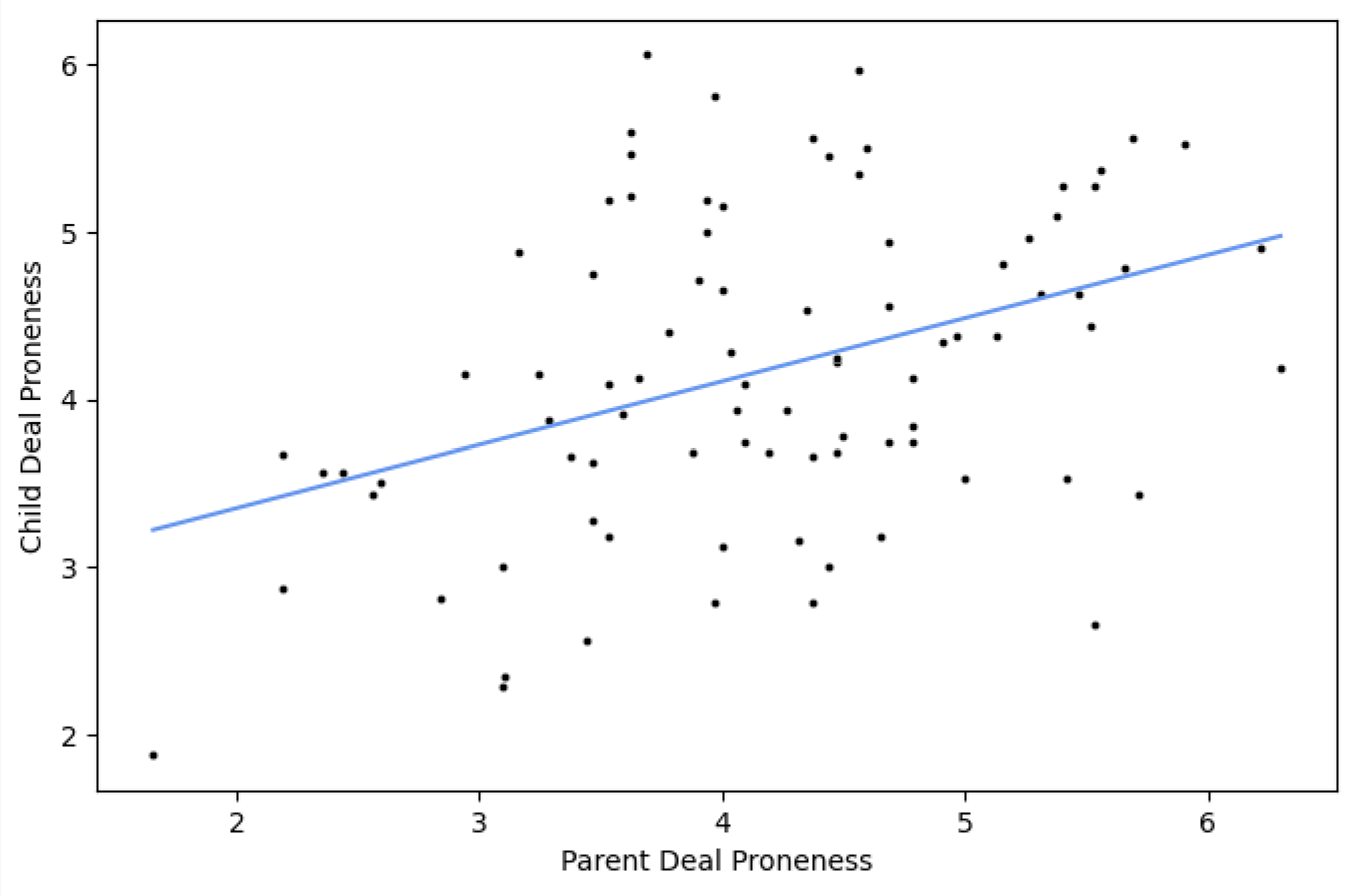


Baseline Model



Credit: Colors selected by Nikhil Lala

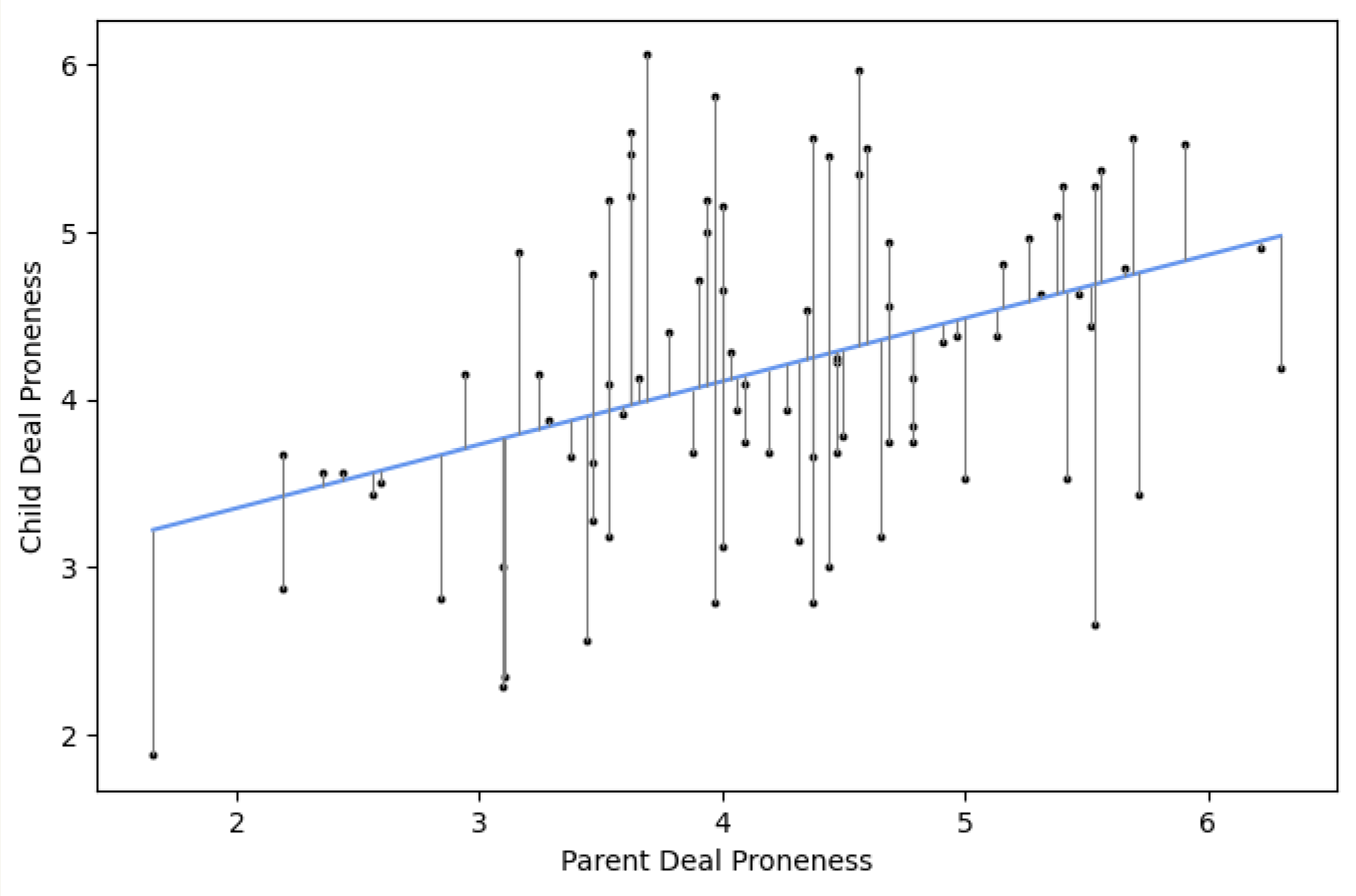
Regression Model



Credit: Colors selected by Rohan Lala

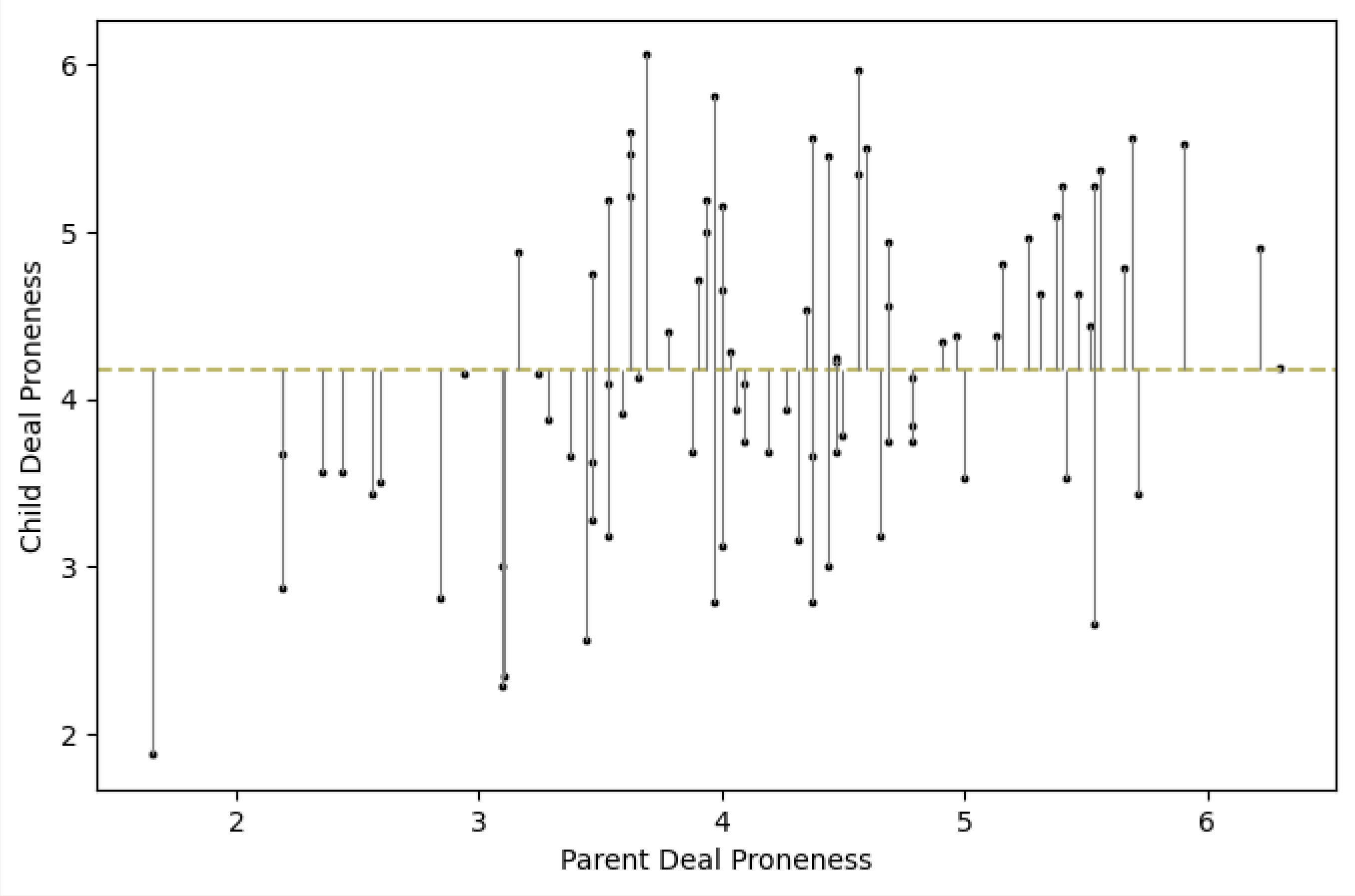
Regression Model (with errors)

sse = min(Σe_i^2) = sum of squared errors



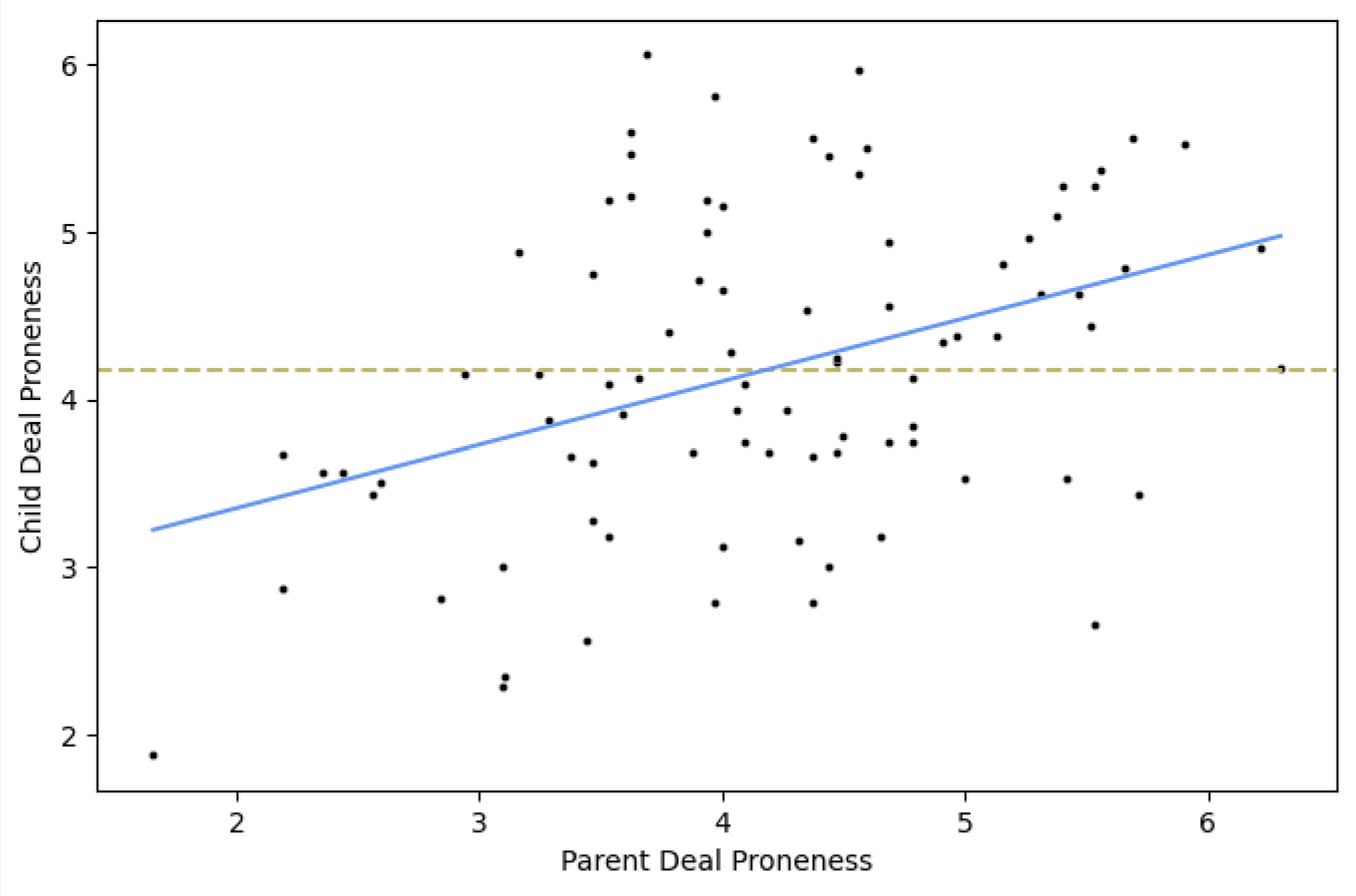
Baseline Model (with errors)

sst = sum of squared total errors



Regression vs. Baseline

$R^2 = 1 - sse/sst$



PREDICTION AND INFERENCE

Prediction

- Is there a relationship between outcome and predictors?
  - *Statistical test to see if at least one of the coefficients is non-zero*
  - *F = ((sst – sse)/p) / (sse/(n-p-1))*
  - *Statistical significance indicates a relationship*
- How strong is the relationship?
  - *R^2 = 1 – sse/sst*
  - *0 < R^2 < 1*
  - *Heuristics: Weak: R^2 <0.1, Moderate: 0.1<= R^2 <0.5; Strong: R^2 >=0.5*
- How accurate are the predictions?
  - *Various indices that incorporate residuals/errors*
  - *Residual error, Sum of squared errors (sse), Mean squared error (mse), Root mean squared error (rmse)*
  - *Cannot be used for comparisons across samples.*

Inference

- Which predictors influence the outcome?
  - *Statistical test to examine individual coefficients*
  - $t = b_1/se(b_1)$; *where* b_1 *is estimate of coefficient for first predictor*
  - *Statistical significance indicates an effect*
- Interpretation of coefficients
  - *A unit change in* X_1 *will result in a change of* b_1 *units in Y while holding all other predictor variables constant.*
- Nature of the relationship (e.g., linear, quadratic, exponential)
  - *Examine scatterplot between predictor and outcome; Statistical significance of non-linear term will reflect nature of relationship.*
- Relative strength of variables
  - *Standardized regression coefficients; Can only be used for predictors in the same model.*
  - *Standardized_b1 = b1*sd(X)/sd(Y)*

Regression Assumptions

- Regression makes a number of assumptions.
- Generally speaking, regression is robust against *small* violations of assumptions.
- It is best to check for these assumptions before conducting analysis.
- A discussion of ways to remedy violations of assumptions is beyond the scope of this course.

- Linear in parameters
- Mean of residuals is zero
- Homoscedasticity
- No autocorrelation
- IVs and residuals are not correlated
- n > number of parameters
- Variance of IVs > 0
- No perfect multicollinearity
- No specification bias
- Errors are normally distributed

MODELS

Simple Regression

- Model relationship between outcome and one predictor
- Numeric Predictor:
  - $y = \beta_0 + \beta_1 x_1 + \varepsilon$
- Categorical Predictor:
  - $y = \beta_0 + \beta_1 x_{1-dummy1} + \varepsilon$
- Categorical Predictor (3-levels):
  - $y = \beta_0 + \beta_1 x_{1-dummy1} + \beta_1 x_{1-dummy2} + \varepsilon$

Simple Regression

- Visualize
  - *Scatterplot (for numeric predictor)*
  - *Bar chart (for categorical predictor)*
- Inference
  - *Statistical significance of coefficient indicates relevance of predictor*
  - *For dummy variables, statistical significance of coefficient indicates difference from reference level*
  - *Value of coefficient indicates the change in outcome for a unit change in predictor*

Multiple Regression

- A multiple regression considers the effects of multiple predictors on the outcome
- Generally speaking, more meaningful predictors will
  - *reduce specification bias by presenting a complete picture (+)*
  - *improve predictions (+)*
  - *lead to overfitting (-)*
  - *reduce interpretability (-)*

Multiple Regression

- Variable Interaction
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$

- Two or more predictors
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_\mathrm{p} x_\mathrm{p} + \varepsilon$

Multiple Regression

- Visualize
  - *Difficult to visualize more than one numeric predictor*
- Inference
  - *Statistical significance of coefficient indicates relevance of predictor*
  - *Value of coefficient indicates the change in outcome for a unit change in predictor (holding all other variables constant)*
  - *Standardized value of coefficients can be used for comparing relative influence*

Compare Models: Out-of-sample

- Model performance is generally,
  - *better on the sample used to train the model*
  - *but worse on data not used to train the model*
- This problem is exacerbated as the model becomes more complex or flexible by say adding more variables.
- Choice of model should be based on comparing candidate models on a test sample

Conclusion

- In this module, we examined
  - *regression*
  - *mechanics of Estimation*
  - *prediction and inference*
  - *types of regression models*