General Strategy for Model Building

David Gerard

2018-12-07

Learning Objective

- Chapter 12
- General Strategy for analysis in multiple linear regression.

Step 1: Identify Objectives and Questions of Interest

- Example 1: Interested in association of one explanatory variable and one response.
- Goal is to determine that association *after adjusting for other variables*.
- Then want to perform variable selection with everything *except* explanatory variable of interest, then include it to test for that association.

Step 1: Identify Objectives and Questions of Interest

- Example 2: Just want to fish for associations
- Then iterate through adding/removing variables, making transformations, checking residuals, until you develop a model with significant terms and no major issues.
- p-values/confidence intervals don't have proper interpretation.
    - Same problems with multiple comparisons — ran many tests and looked at data a lot to come to final model.
- You generally build a model and tell stories with it.

Step 1: Identify Objectives and Questions of Interest

- Example 3: Prediction
- Include variables to maximize predictive power, don't worry about interpretation.

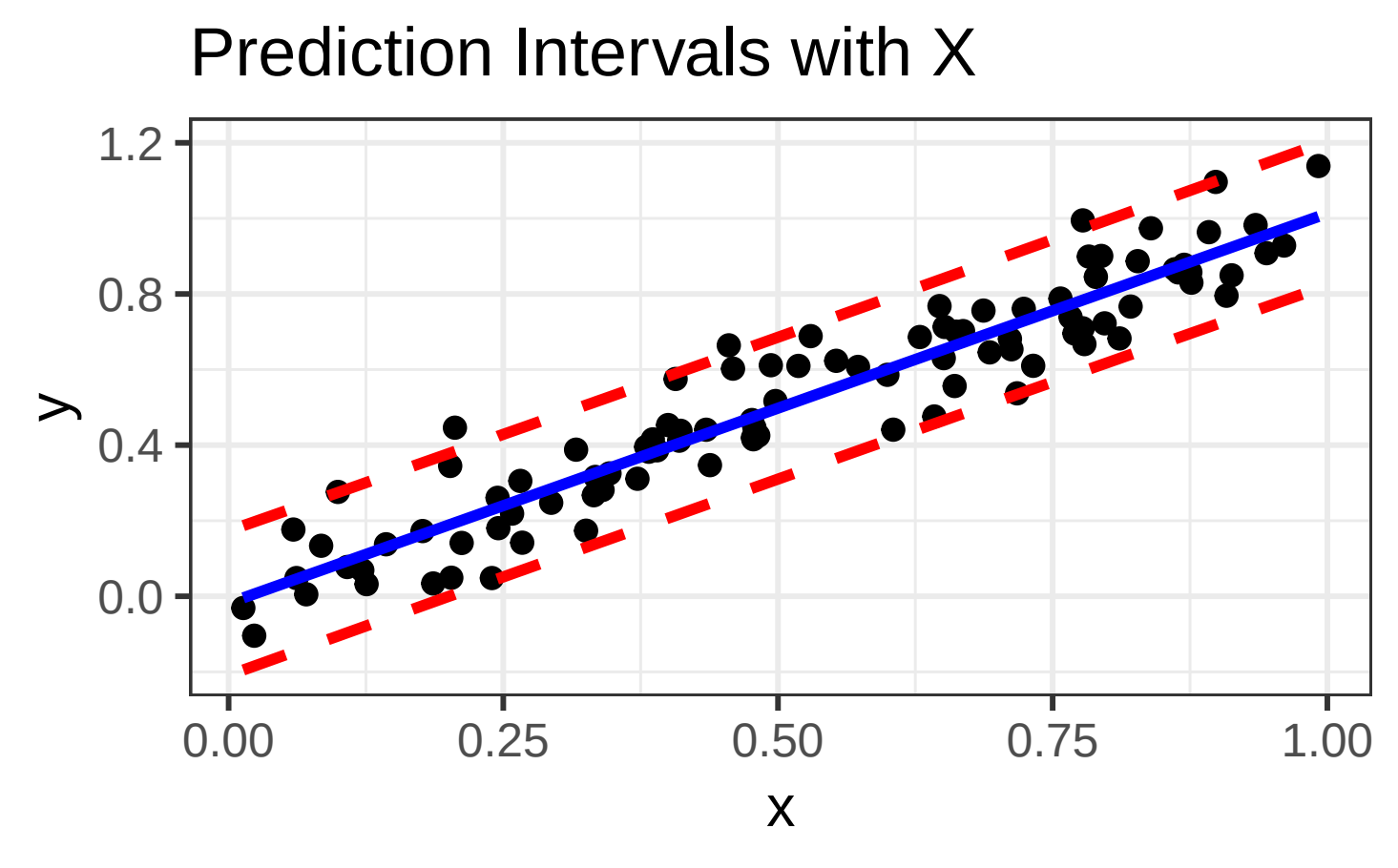
Step 2: Screen Available Variables

- Choose a list of explanatory variables that are important to the objective.
- Screen out redundant variables

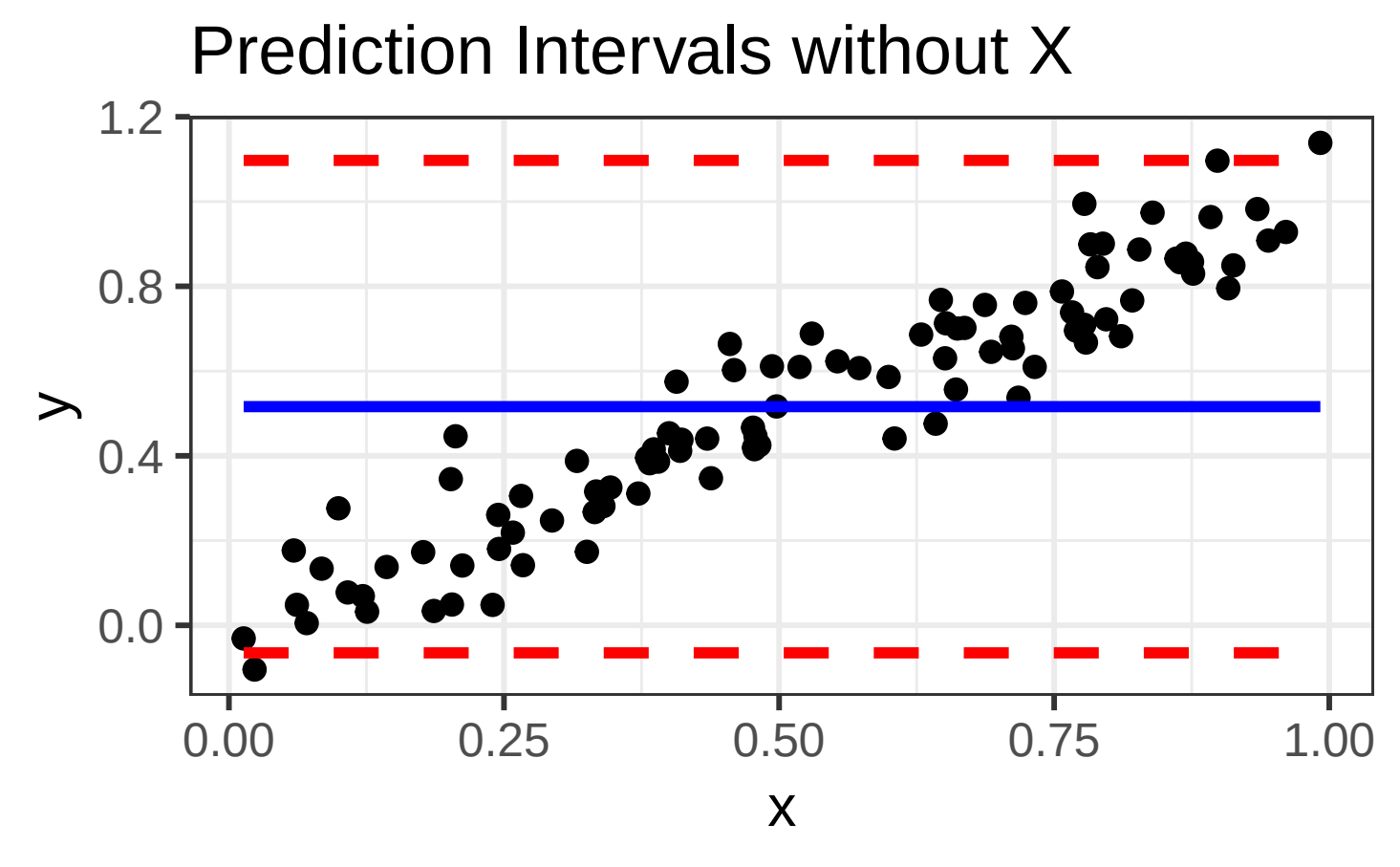
Problems with Including Too Few Variables

- You are only picking up **marginal** associations.
- E.g., we already know that men make more money than women. We want to see if men **still** make more money than women when we control for other variables.
- Predictions are less accurate.

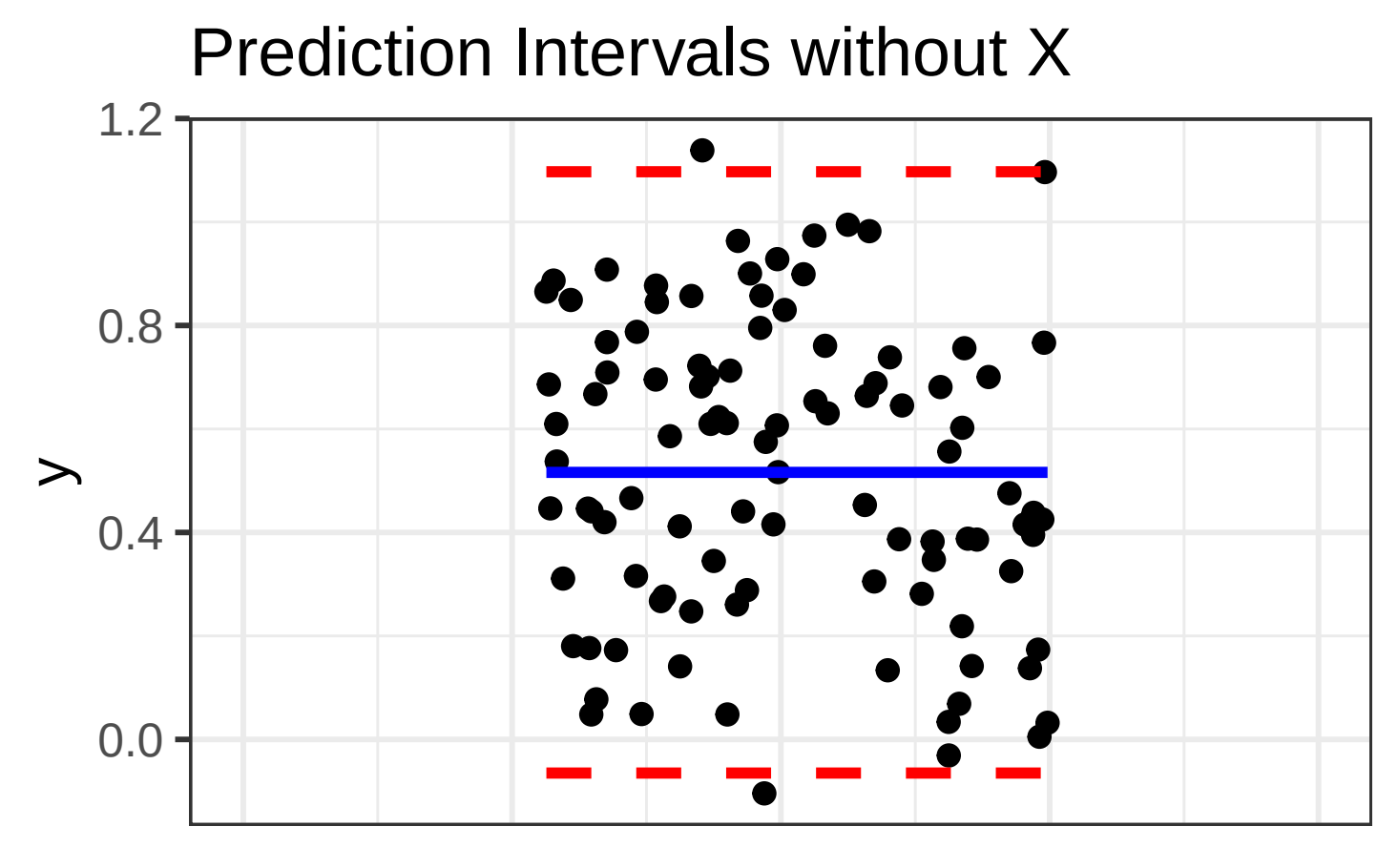
Too few variables: Predictions are less accurate



Too few variables: Predictions are less accurate



Too few variables: Predictions are less accurate



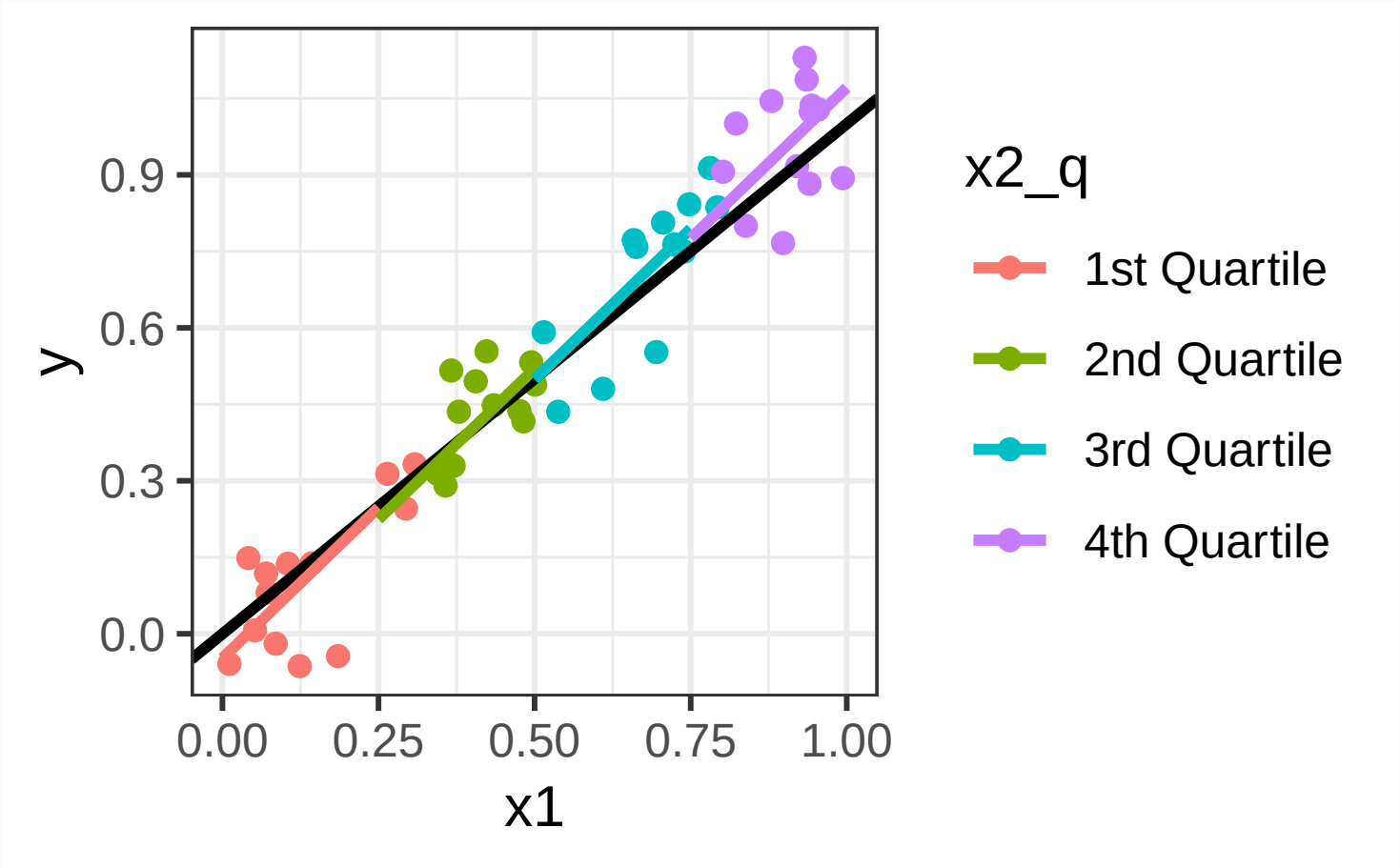
Problems with too many variables

- Harder to estimate more parameters.
- Formally, the variances of the sampling distributions of the coefficients in the model will get much larger.
- Including highly correlated explanatory variables will **really** increase the variance of the sampling distributions of the coefficient estimates.
- Intuitively, we are less sure if the association of Y and X_1 is due to that actual associate or is it mediated through X_2?
- Predictions are less accurate.

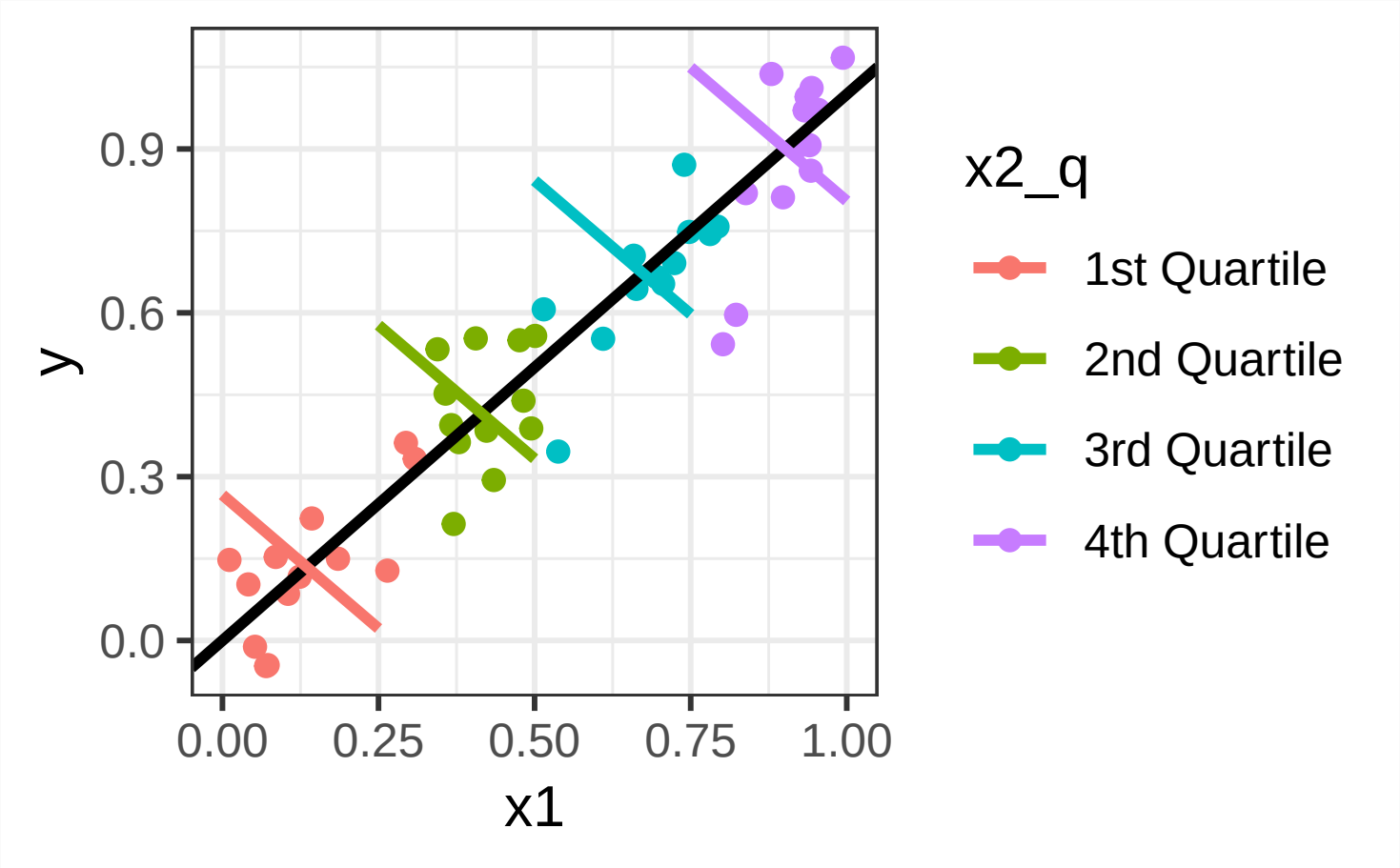
Demonstration

- True model: $\mu(Y|X_1) = X_1$
- Fit Model: $\mu(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- Correlation between X_1 and X_2 is 0.9994.
- We will simulate Y and plot the resulting OLS estimates.

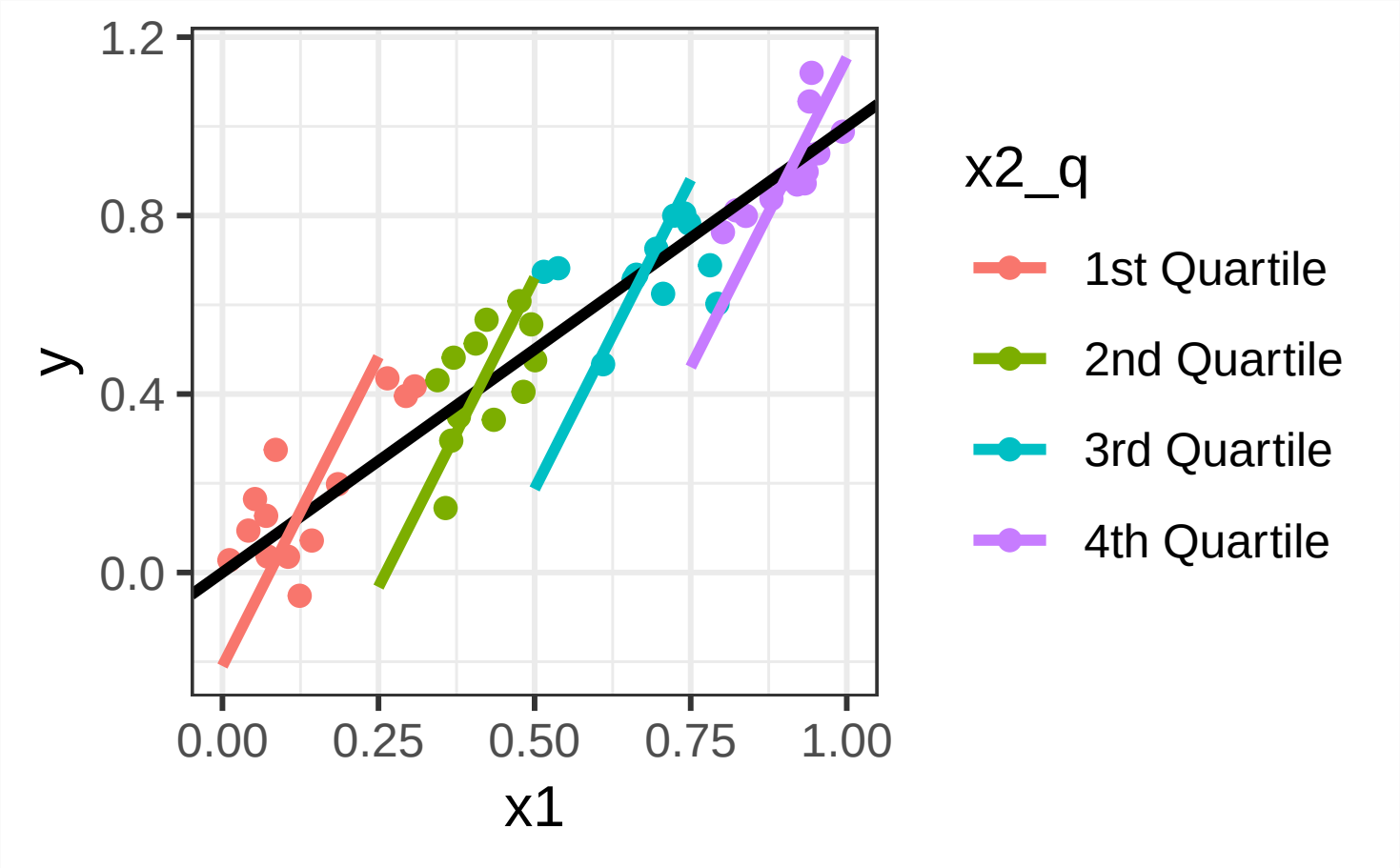
Demonstration: Black is truth



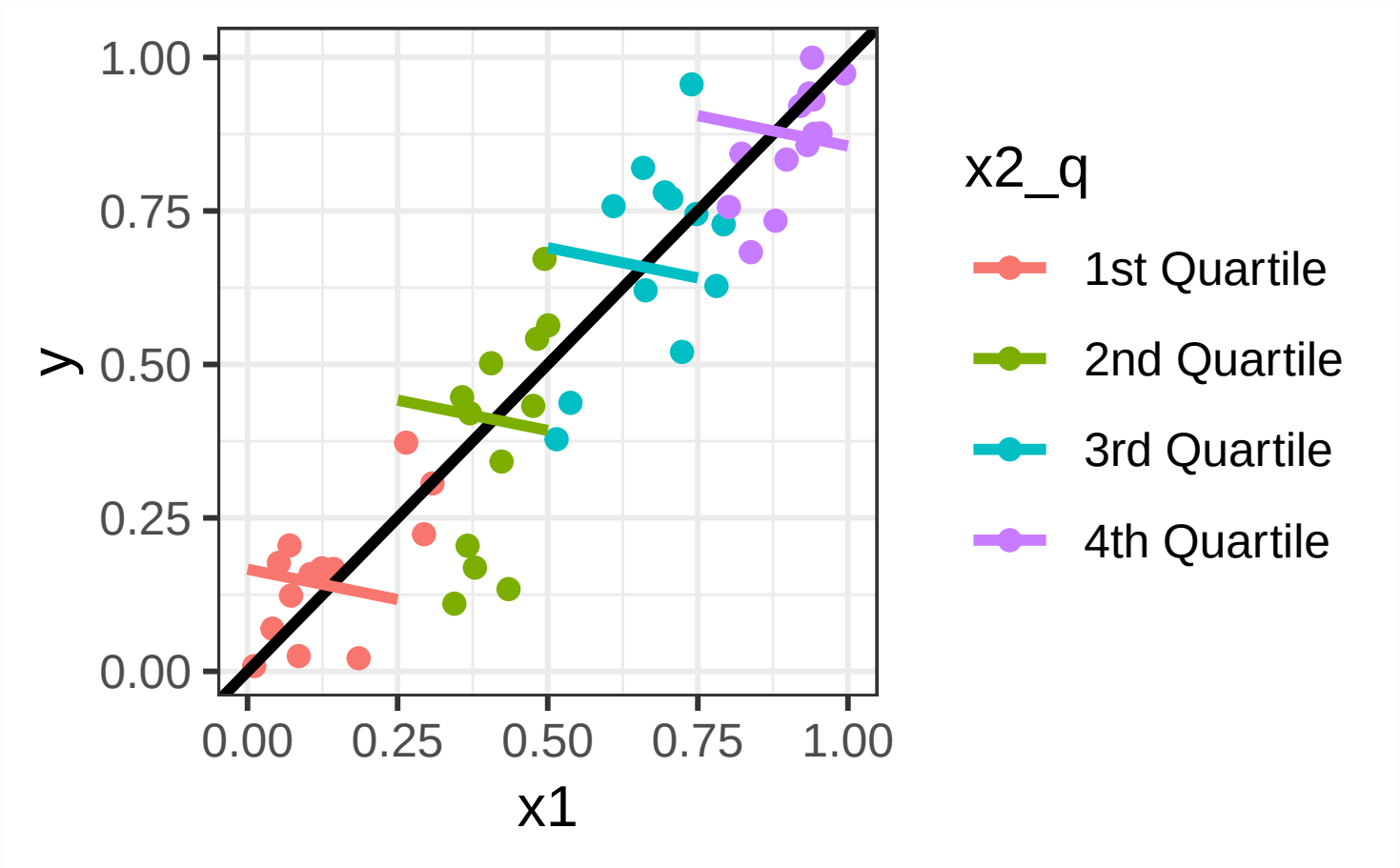
Demonstration: Black is truth



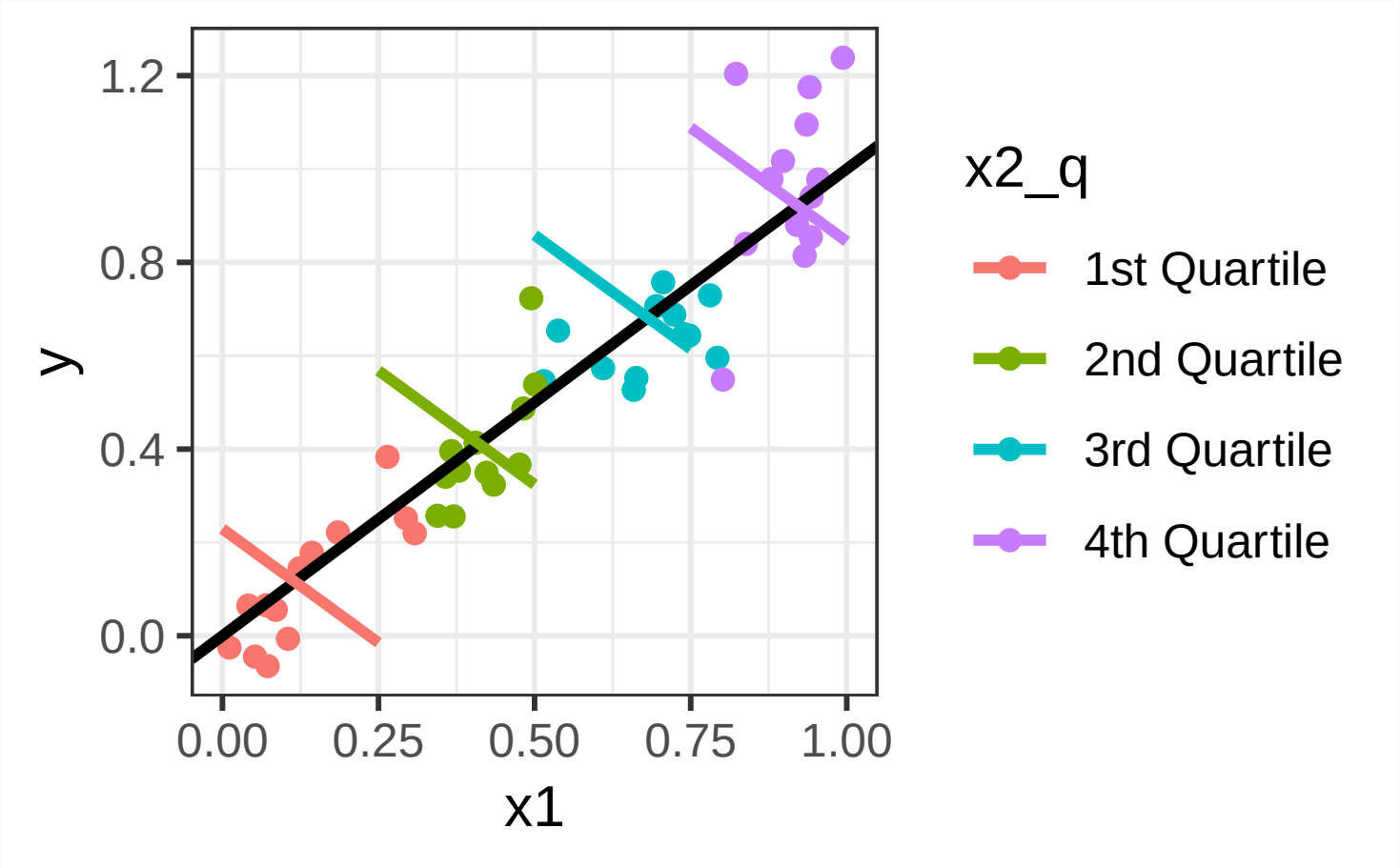
Demonstration: Black is truth



Demonstration: Black is truth



Demonstration: Black is truth



Steps 3 through 5

- Exploratory data analysis.
  - Tons of scatterplots.
  - Look at correlation coefficients.
  - 12_multiple_regression_eda.pdf
- Transformations based on EDA.
  - 11_linear_model_assumptions.pdf, 11_interpreting_log_transformations.pdf
- Fit a rich model and look at residuals.
  - Look for curvature, non-constant variance, and outliers.
  - 14_outlier.pdf, 11_linear_model_assumptions.pdf, 12_multiple_regression_eda.pdf
- Iterate the above steps until you don't see any issues.

Step 6

- If appropriate, use a computer aided technique to choose a suitable subset of explanatory variables.
  - 13_f_test_of_nested_models.pdf
  - 13_non_nested_comparisons.pdf

Step 7

- Proceed with analysis with chosen explanatory variables.
- Tell stories with the data using p-values, coefficient estimates, confidence intervals, etc. . .