Covariate Shift and Confounding

27 November 2016

Common Assumptions of Supervised Learning

- Joint distribution of (X,Y) that produces the training data is the same is used to evaluate generalization error.

- By axioms of probability,
  $P(X=x,Y=y) = P(X=x)*P(Y=y|X=x)$

- So we can decompose into two assumptions:

  - Distribution of X stays the same from training to generalization

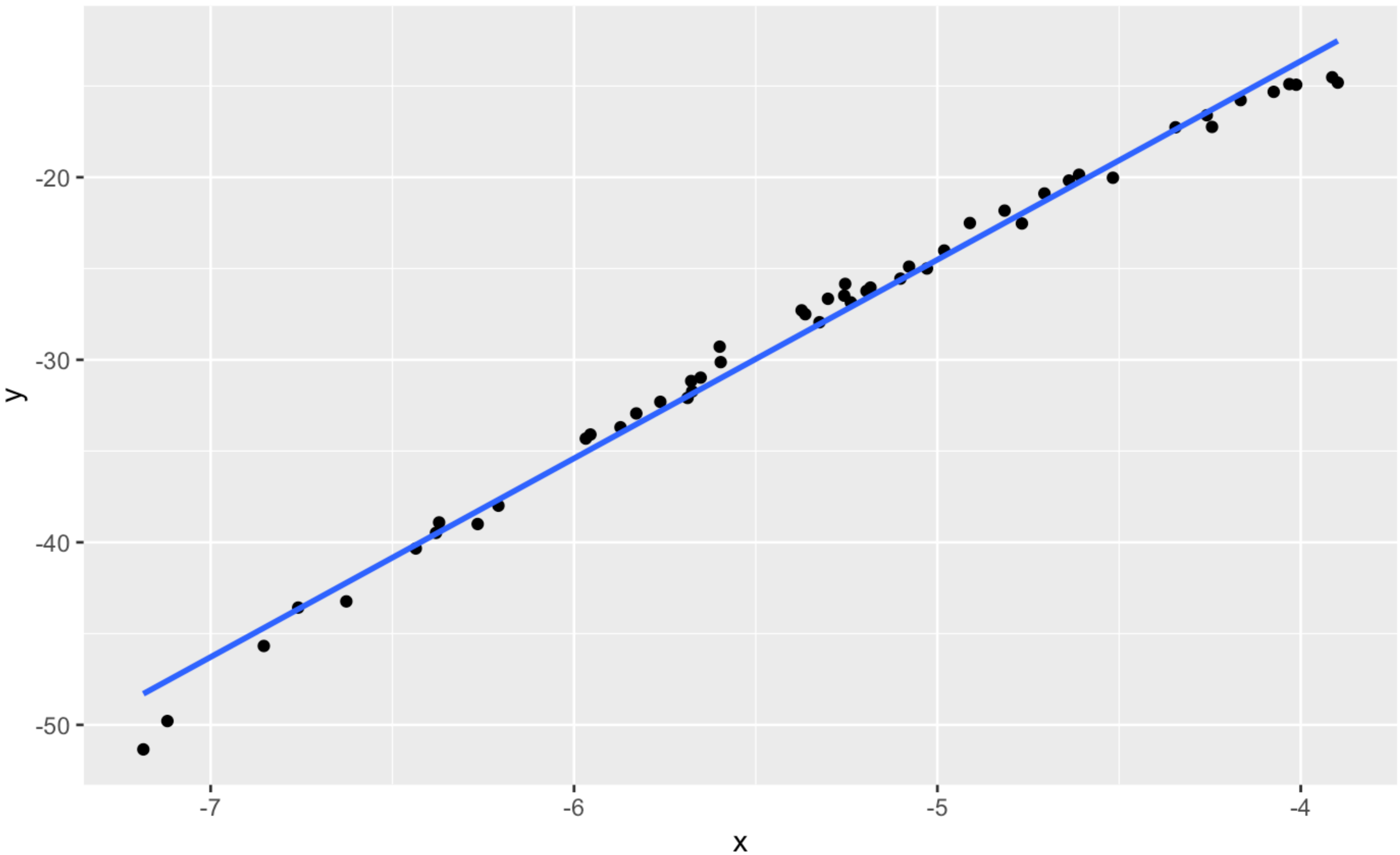
  - Distribution of Y|X stays the same from training to generalization

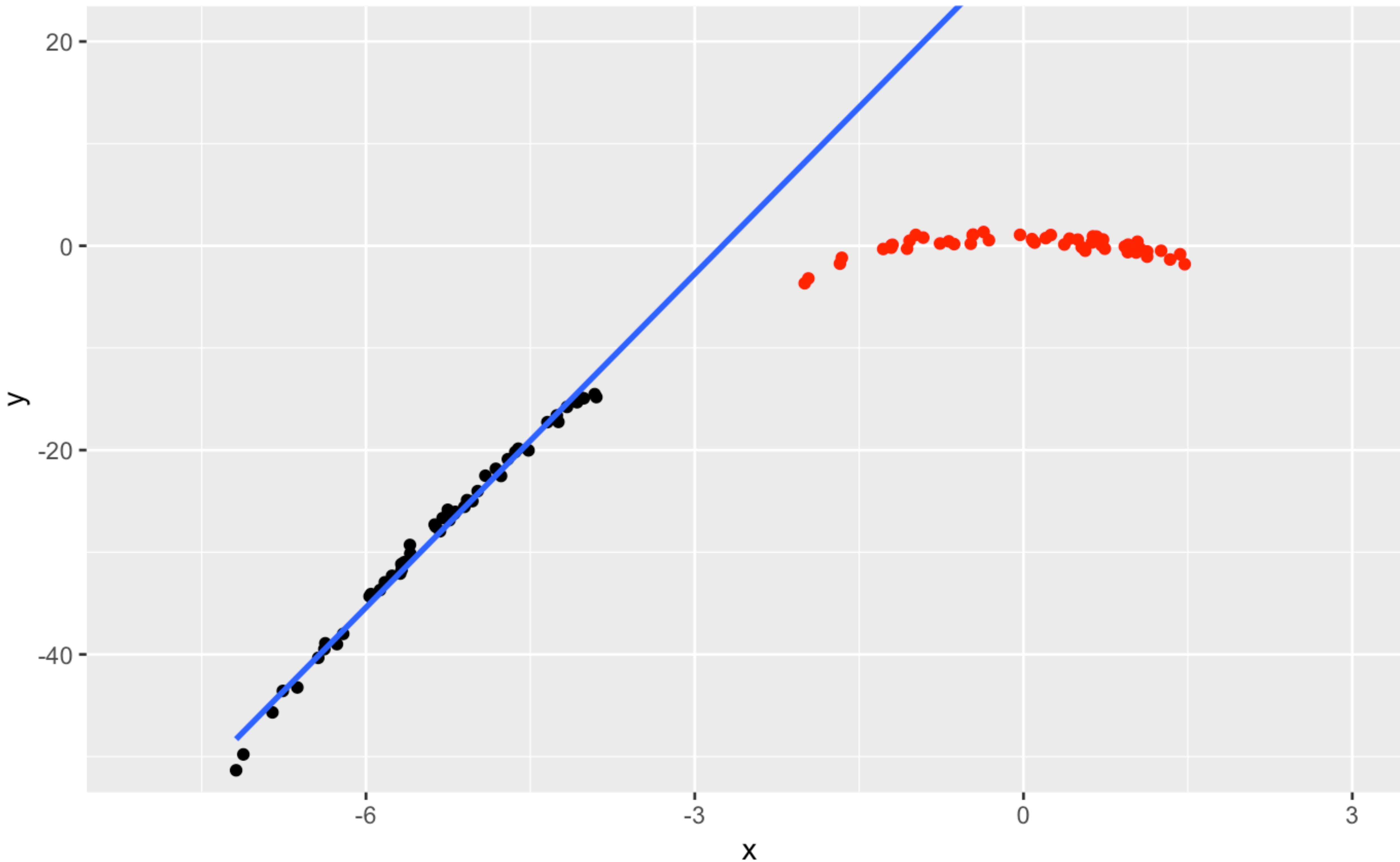
What if P(X) changes?

- Lets see what can happen if the **marginal distribution of P(X = x) changes from training to "future-test."** (Imagine a *new* test set that is given that was not available during training. We will call this the "future-test" set.)

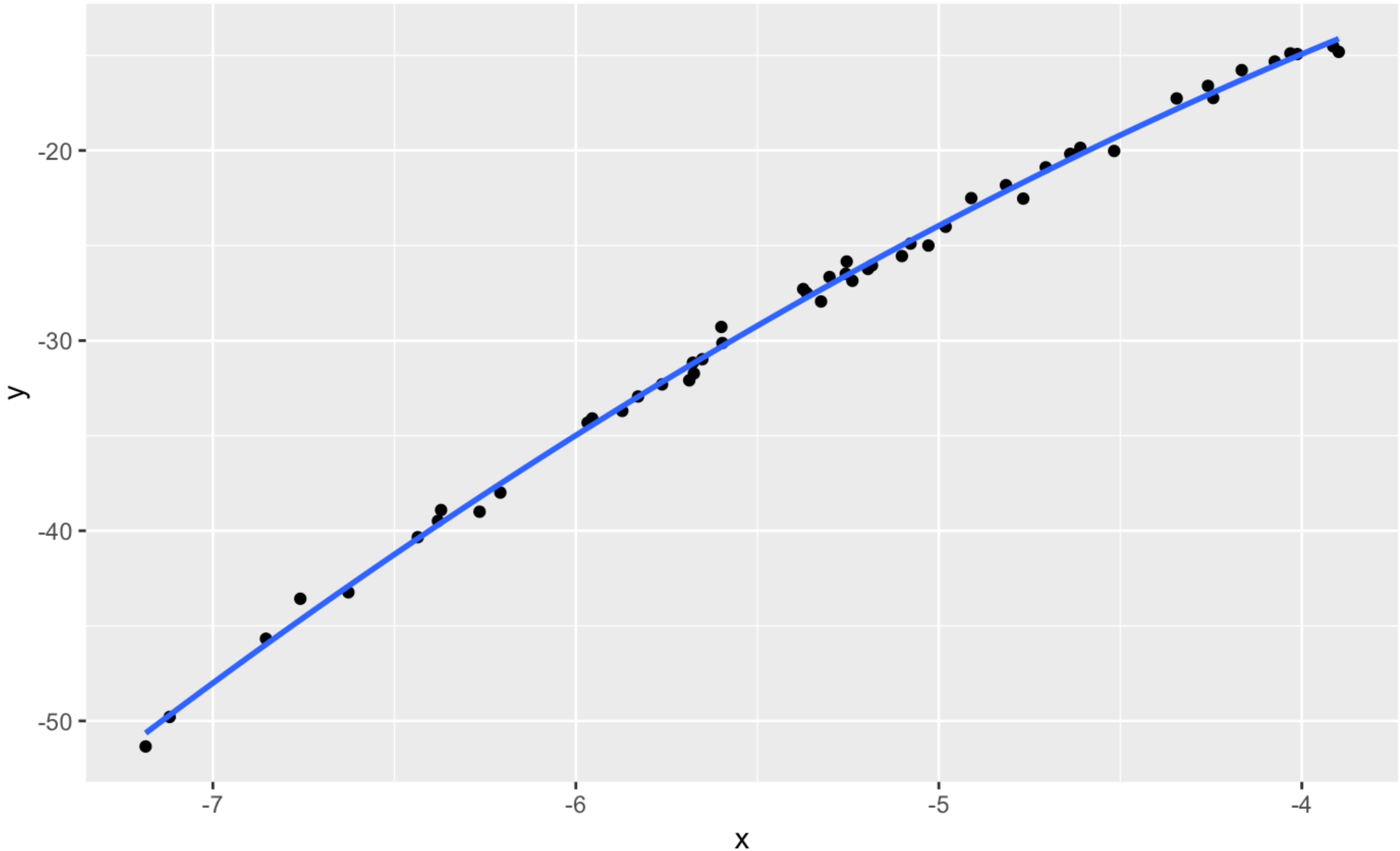
- This is called **covariate shift**

- In these examples, we will make sure that Y|X does not change - the true relationship between X and Y stays the same in all examples.

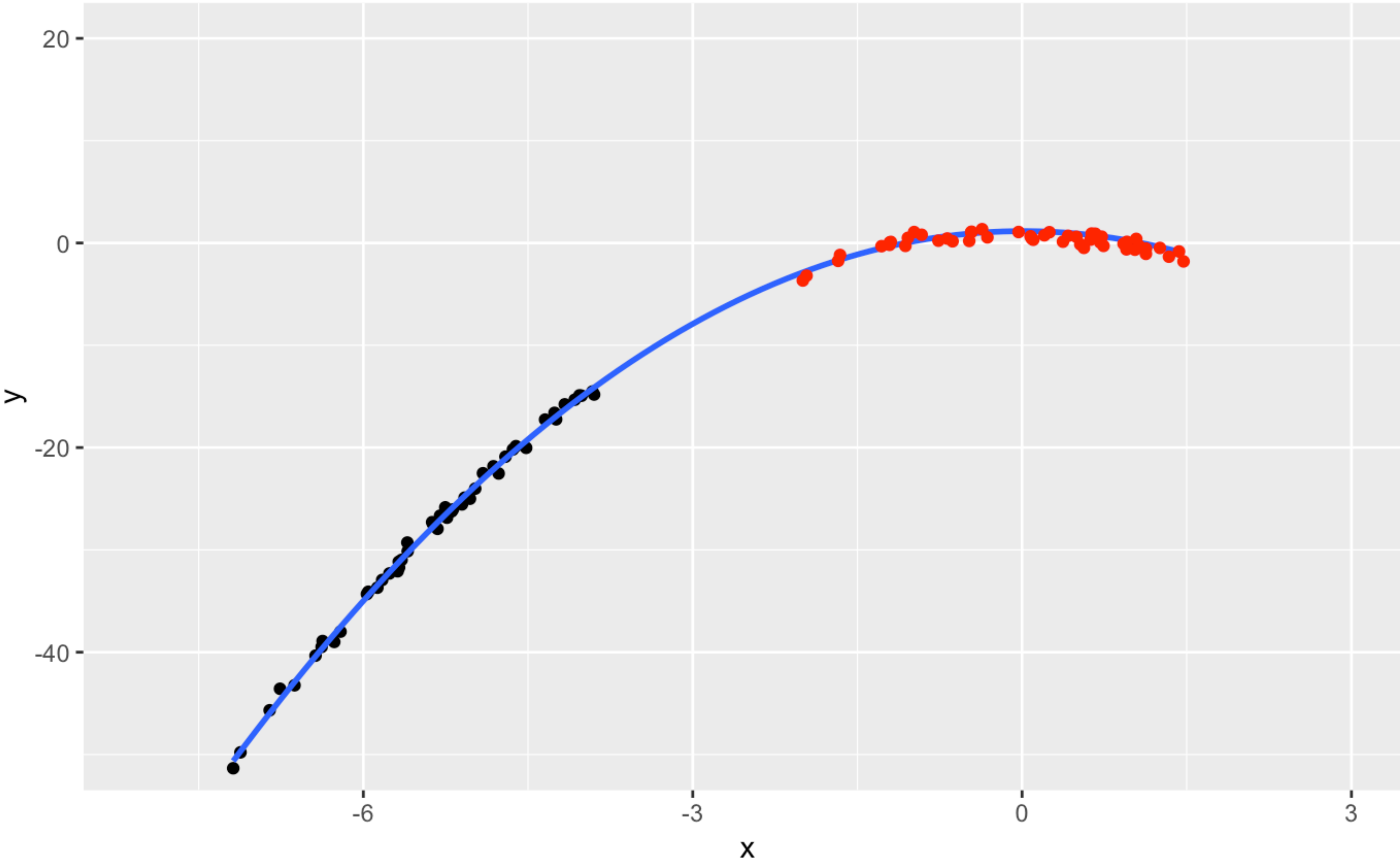

-20
-30
-40
-50
y
-7
-6
-5
-4
x



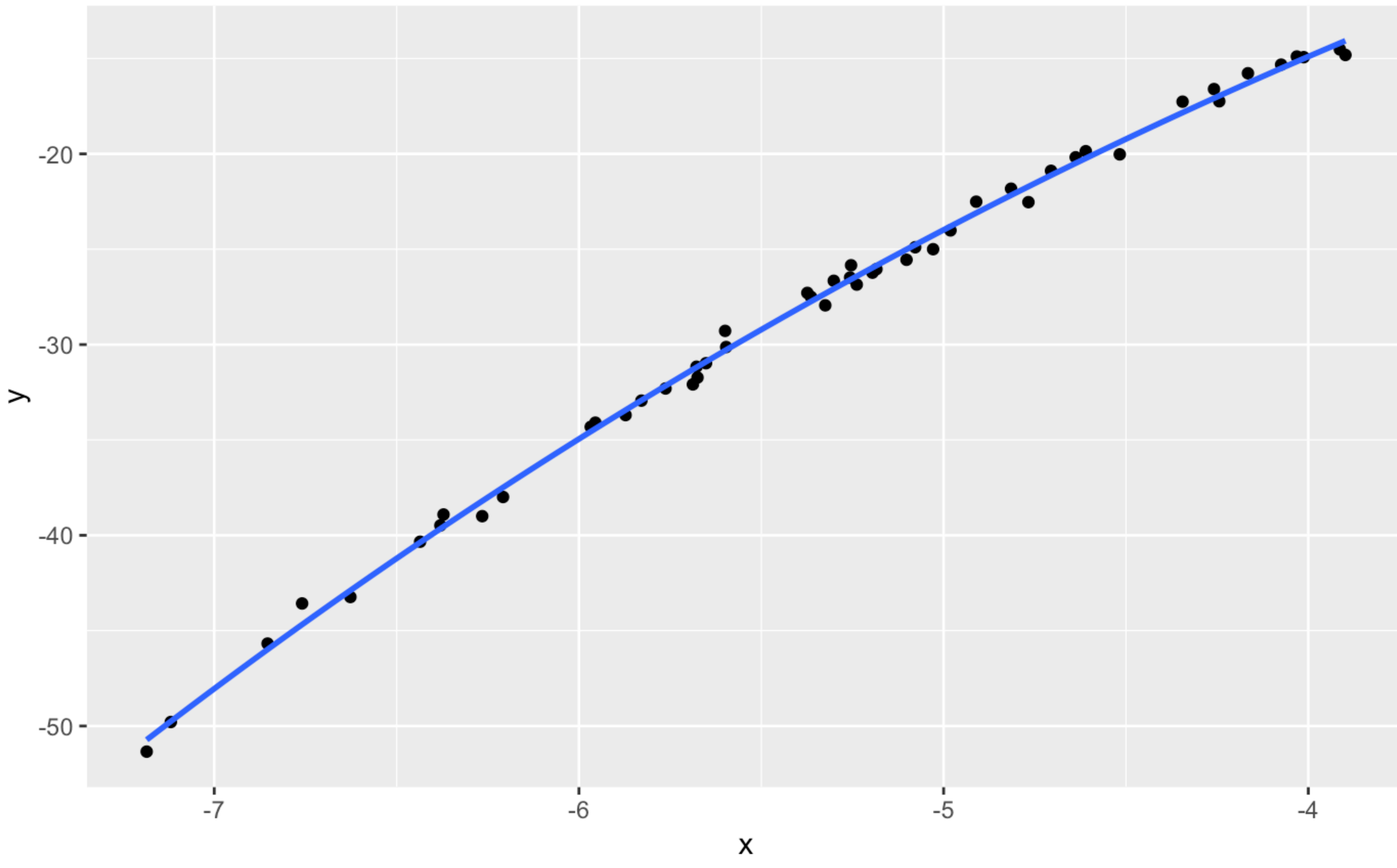
True model is quadratic; underfitting with linear model



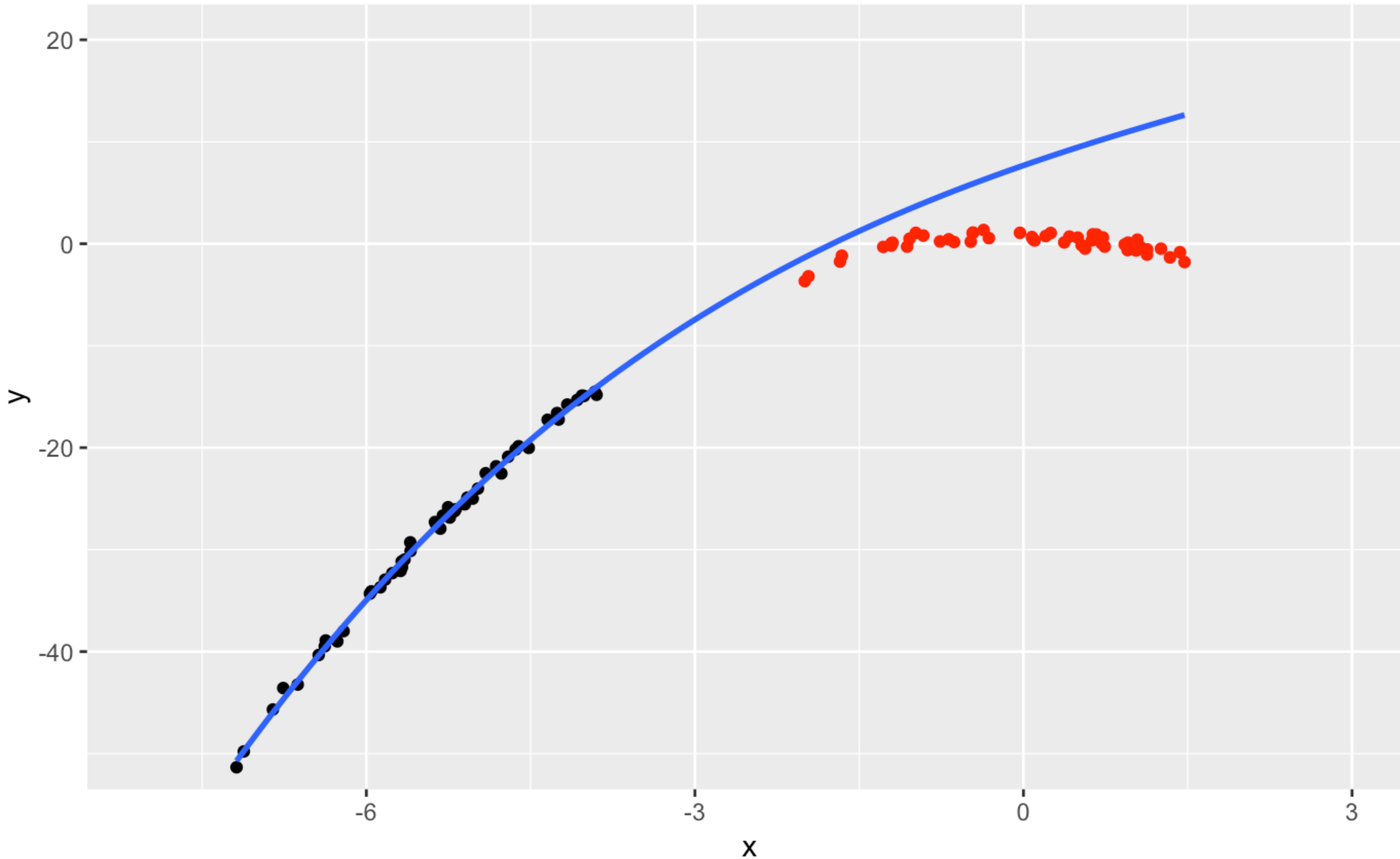
-20
-30
-40
-50
y
-7
-6
-5
-4
x



True model is quadratic, correctly captures relationship

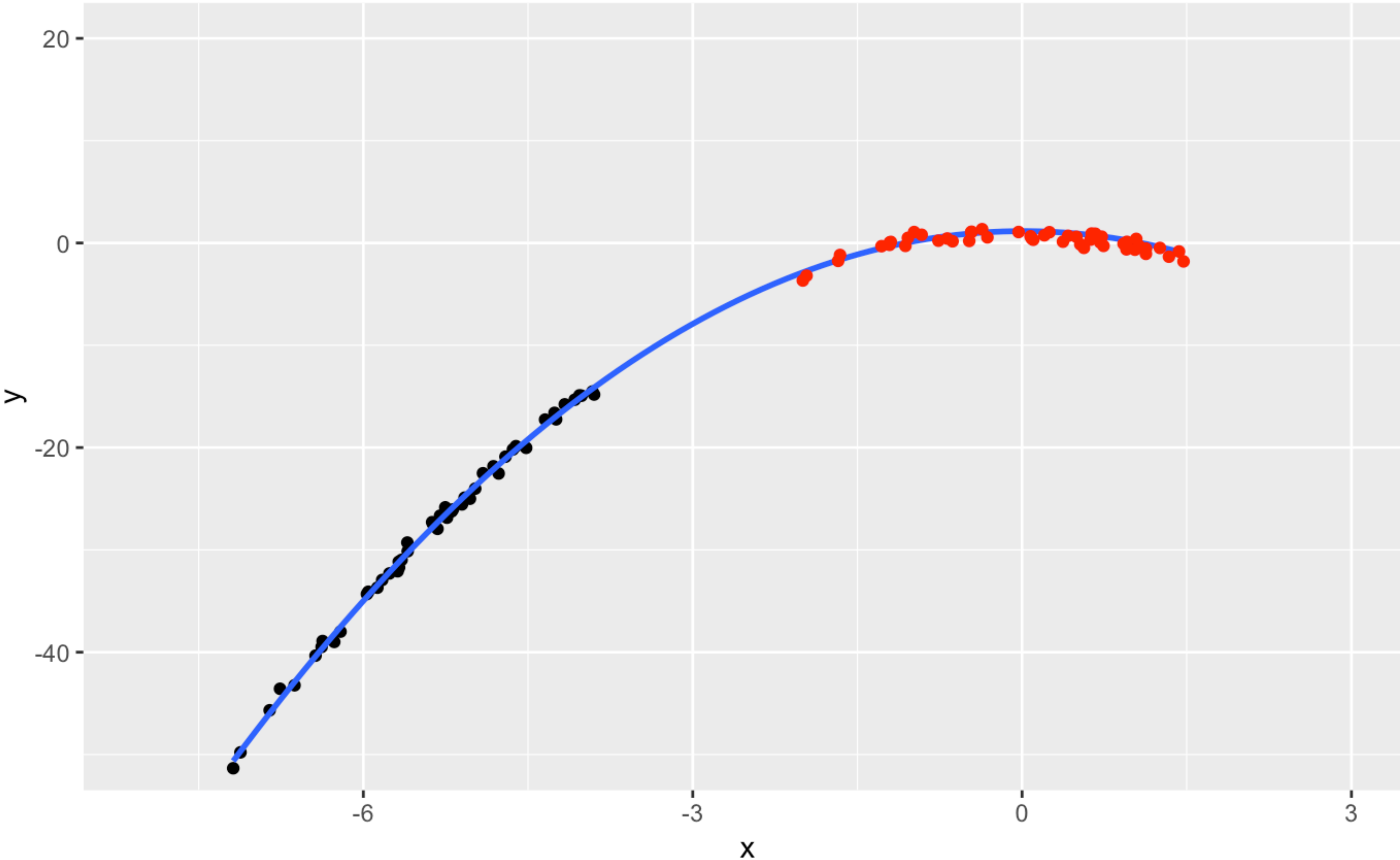


-20
-30
-40
-50
y
-7
-6
-5
-4
x

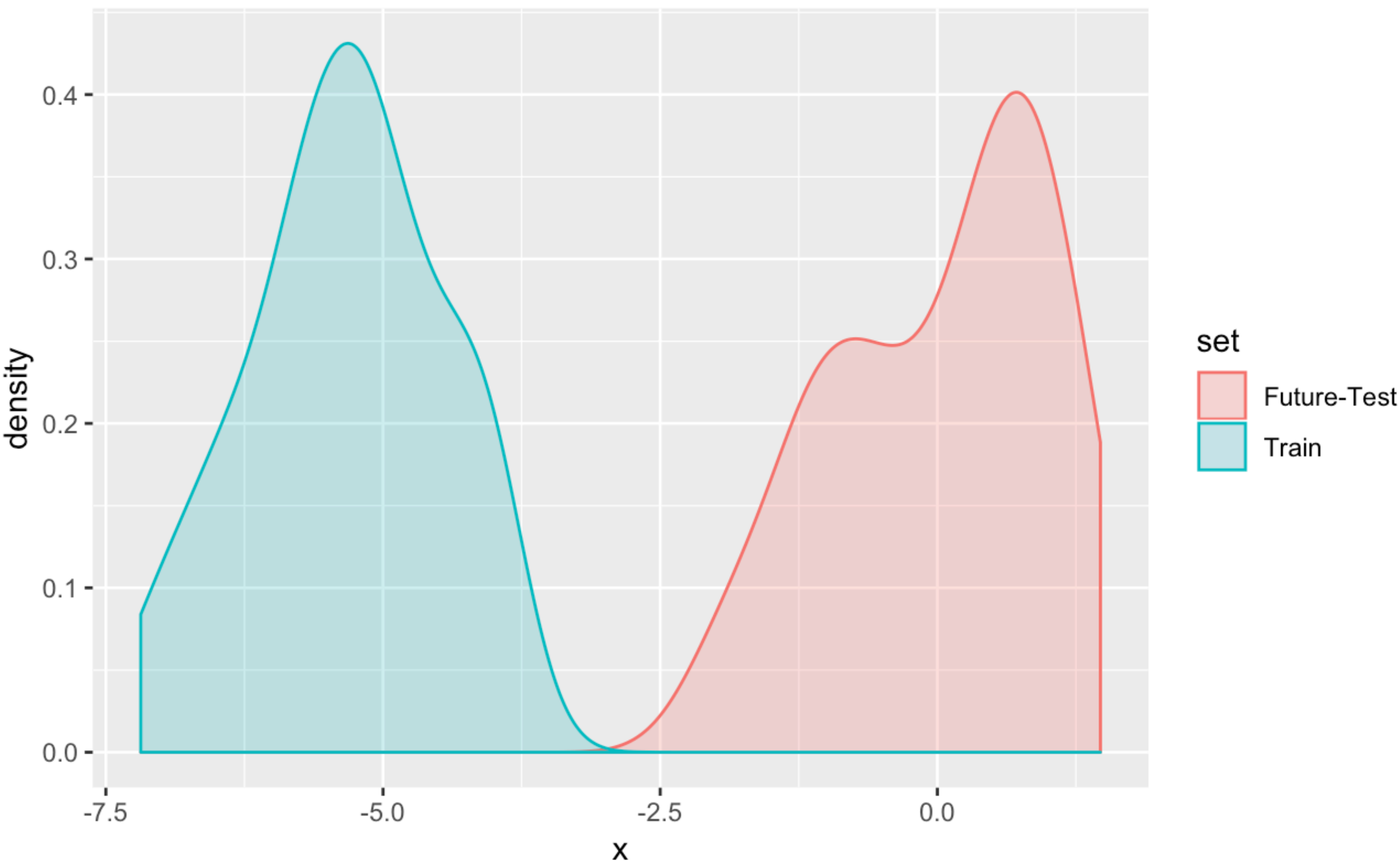


True model is quadratic — overfitting with cubic model

Can we detect covariate shift?



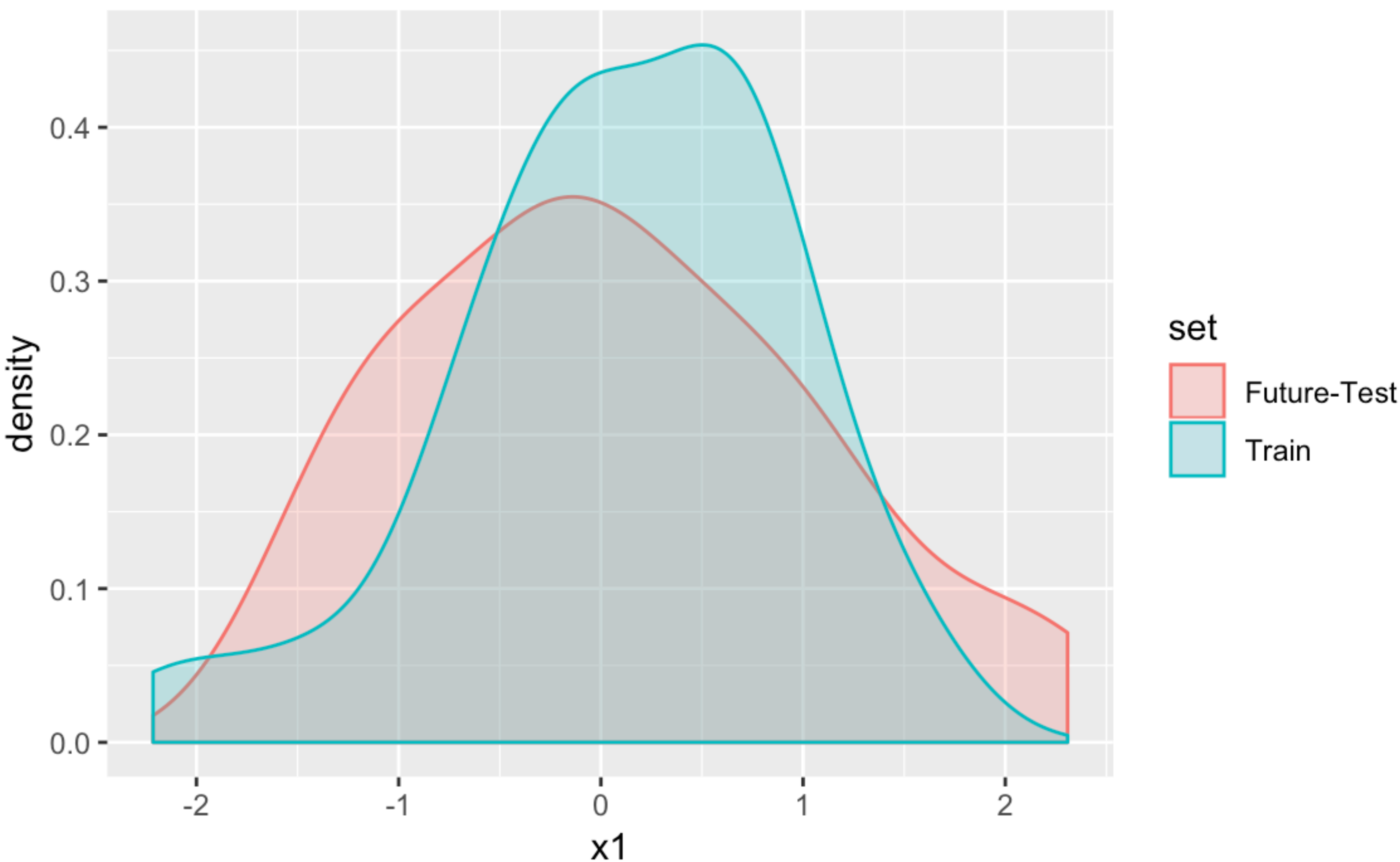
True model is quadratic, correctly captures relationship



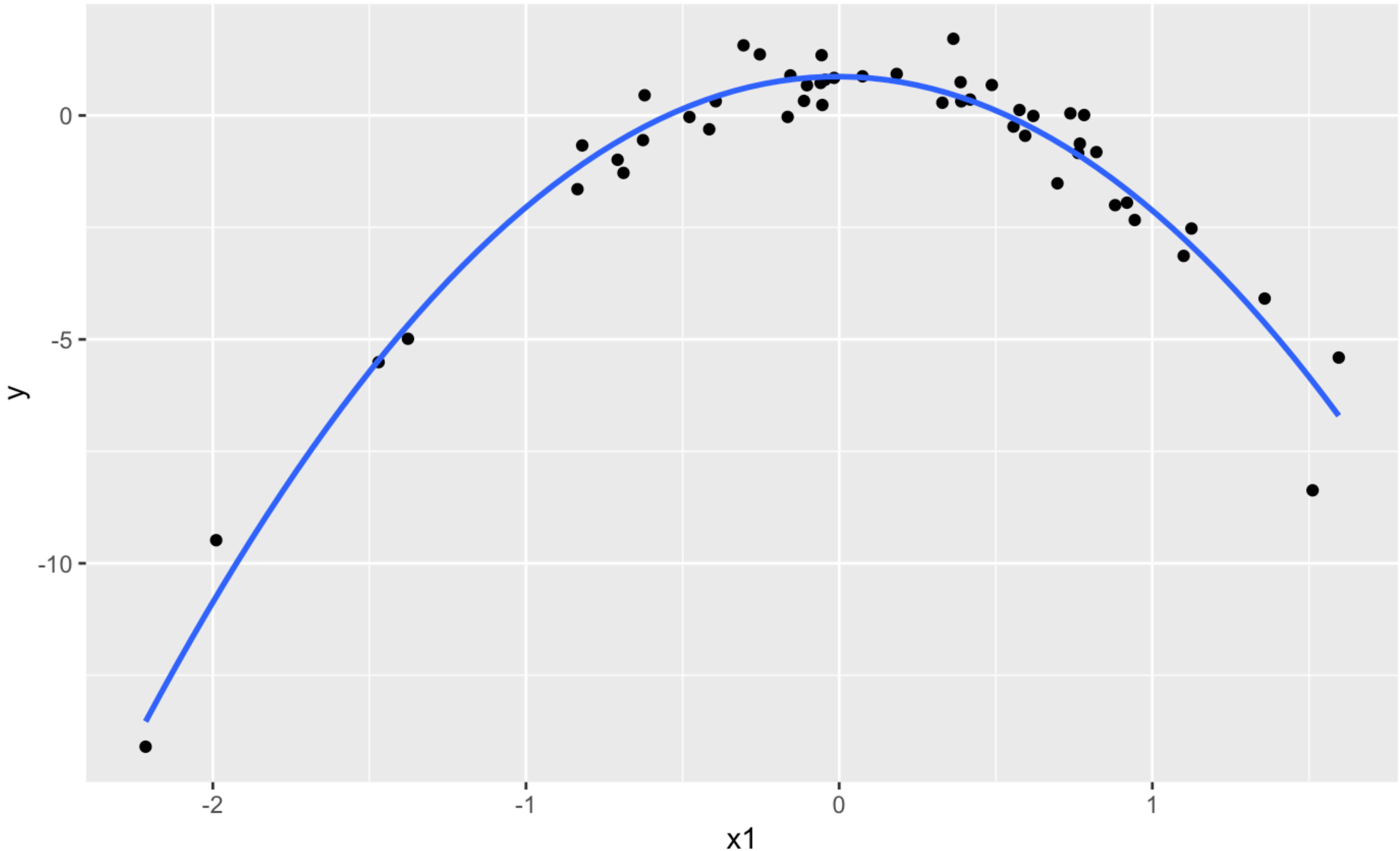
density
0.0
0.1
0.2
0.3
0.4
-7.5
-5.0
-2.5
0.0
x
set
Future-Test
Train

Detecting covariate shift

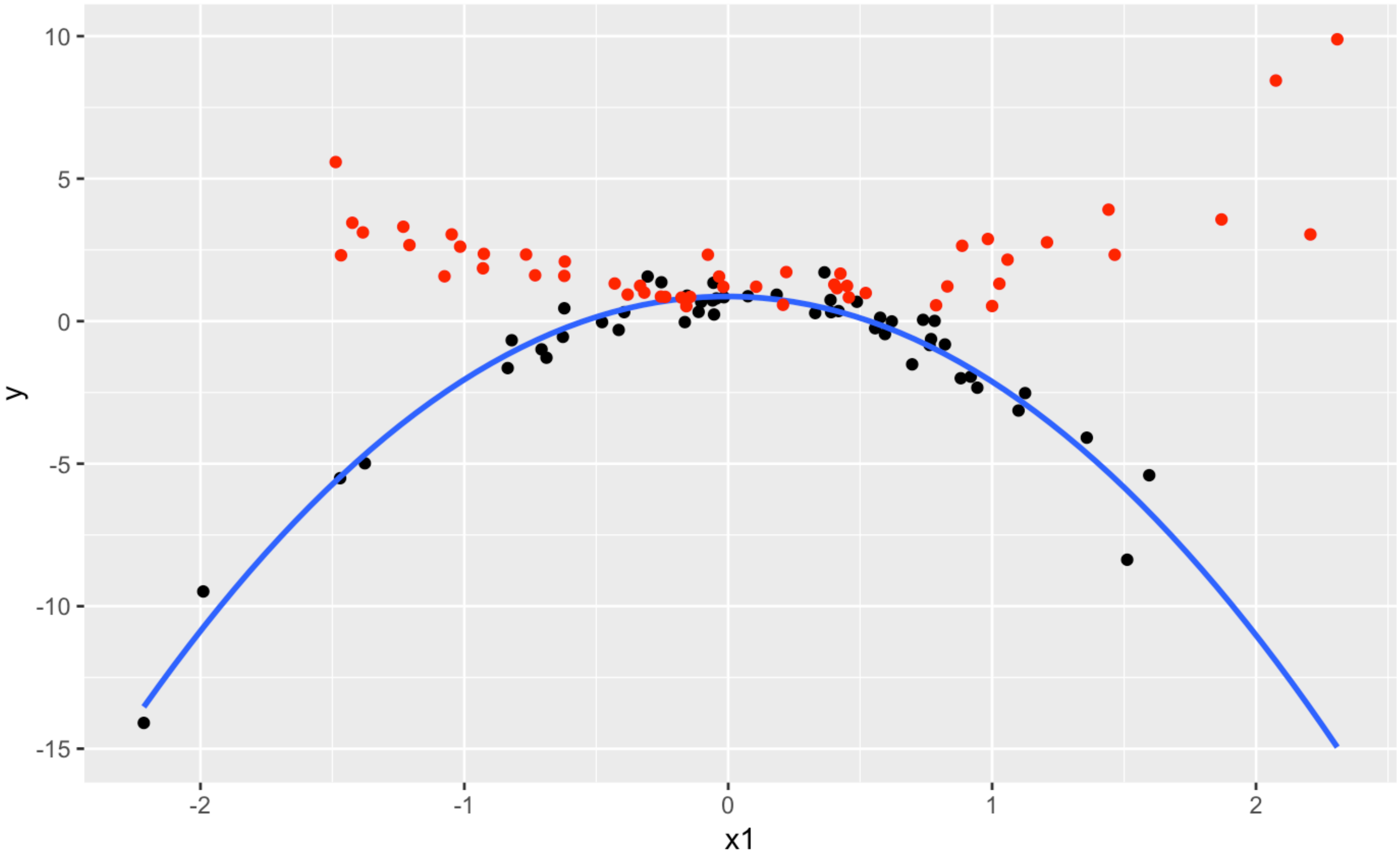
- **Train a classifier to distinguish training instances from future-test instances**
- If one can be learned that gives good performance, probably a covariate shift has occurred
- This may have limited usefulness in practice; might be as difficult as just learning a new model for the original task.



density
0.4
0.3
0.2
0.1
0.0
-2
-1
0
1
2
x1
set
Future-Test
Train

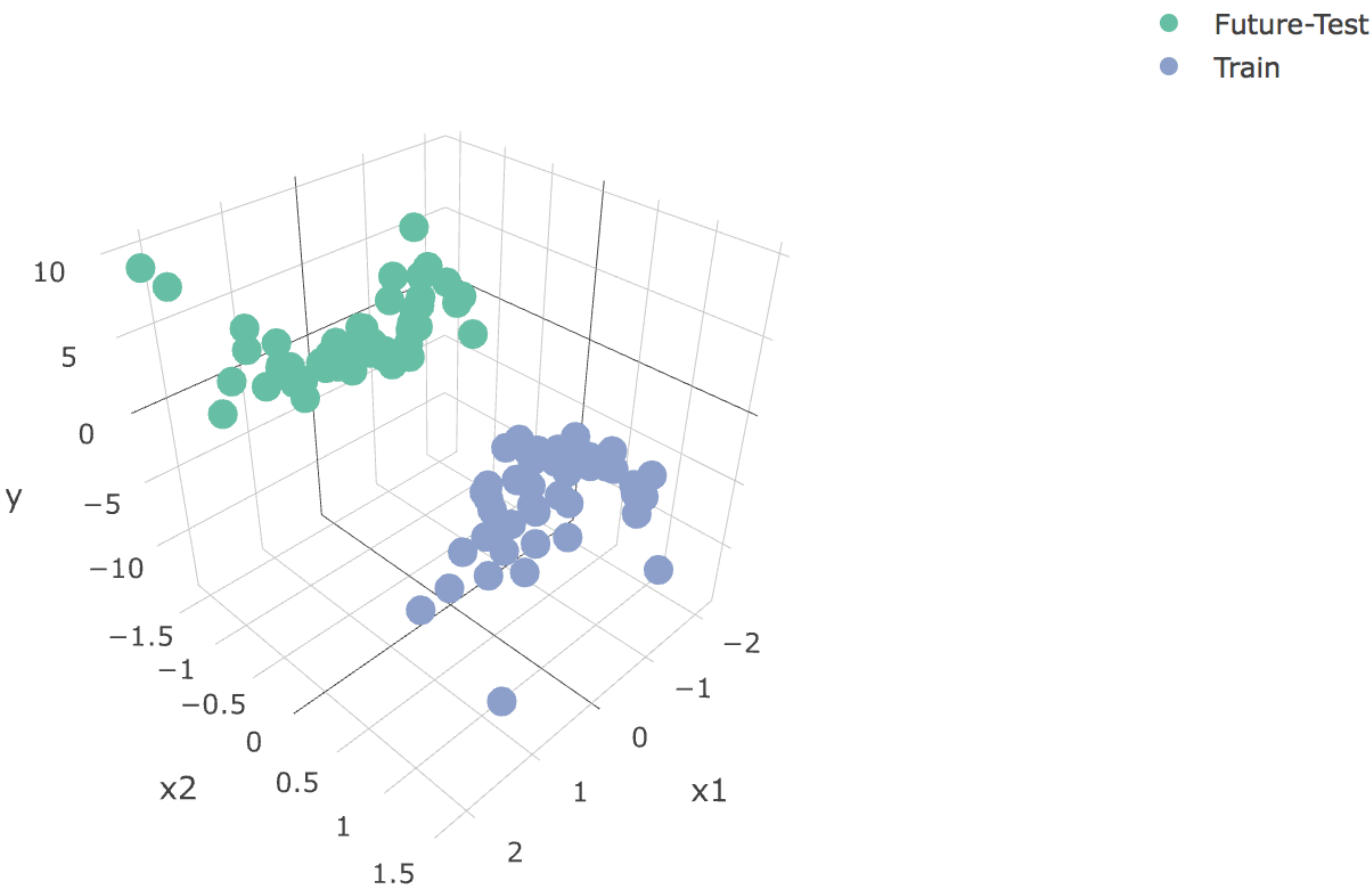


y
0
-5
-10
-2
-1
0
1
x1



10
5
0
-5
-10
-15
y
-2
-1
0
1
2
x1

Has the relationship between y and x changed?


Future-Test
Train
y
10
5
0
−5
−10
x2
−1.5
−1
−0.5
0
0.5
1
1.5
x1
−2
−1
0
1
2

Omitted Variables

- Omitted variables can cause significant problems if they undergo covariate shift from training to future-test.

- They can also cause significant problems with model interpretation.

Confounding

- In statistics, **a confounder (also confounding variable, confounding factor or lurking variable) is a variable that influences both the dependent variable (e.g. class label) and one or more independent variables (e.g. features) causing spurious associations.**

- When we have *unmeasured confounders* (confounding variables omitted from the model), we can get misleading results.

- Confounding is a *causal* concept; it cannot be addressed with statistics alone.

- This situation is sometimes called *endogeneity* in economics.

Example: Medication Effectiveness

- Consider a very simple dataset with one feature UseMed, and a label Cured.

- Model 1: Logistic regression
  - P(Cured = 1 | UseMed) = s(-0.06014 + 2.03604*UseMed)
  - AUC = 0.72

Example: Medication Effectiveness

- Consider a very simple dataset with a feature UseMed, a feature to tell if a person is Rich, and a label Cured.

- Model 2: Logistic regression

  - P(Cured = 1 | UseMed, Rich) = s(-0.58 - 0.10*UseMed + 2.96*Rich)

  - AUC = 0.82

Example: Medication Effectiveness

- Consider a very simple dataset with a feature UseMed, a feature to tell if a person is Rich, a feature to tell if a person is Healthy, and a label Cured.

- Model 3: Logistic regression

  - P(Cured = 1 | UseMed, Rich, Healthy) = s(-1.26 - 0.12*UseMed + 0.38*Rich + 3.60*Healthy)

  - AUC = 0.90

Confounding in Practice

- Confounding makes interpreting models challenging.
- Always important to consider what other — possibly unmeasured — features could influence outcome/label.
- ***There is no statistical test for confounding. Test and validation sets, cross-validation, bootstrapping, etc. cannot detect confounding.***

Bias in ML More Generally

- Supervised learning methods learn whatever patterns in the training data are most useful for prediction.
- Sometimes, this perpetuates undesirable patterns.
- https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
- http://science.sciencemag.org/content/356/6334/133