
# PPOL 502-07: Reg. Methods for Policy Analysis

## Week 12: Predictive Models

Alexander Podkul, PhD
Spring 2022

# Today's Class Outline

- Course Schedule
- Predictions (Review)
- Prediction Errors
- Using Predictions
- Working with Out of Sample Data
- Fitting the Model
- Where We're Going Next
- **Break**
- Working in Stata

# Course Schedule

### Tonight (4/13)

- {Nothing due}

### Next week (4/20)

- Data Project due

### The final week (4/27)

- {Nothing due}
- Post suggested readings! (or email them to me)

### The final exam (5/6)

- Final exam!

# Data Project Comments

General update:

- (Unless you have an extenuating circumstance) Feedback has been distributed!
- Projects seem to be in a good place!

A few focus areas as we approach the final stretch:

- Consider multiple specifications (telling the whole story)
- Think hard about omitted variable bias
- Perform diagnostics
- Transforming variables may add to your story
- It's never too late to consider interaction or quadratic models 😄

# Reviewing Predictions

So far this semester, we've already discussed a number of prediction-related topics:

- Fitted Values
- Adapting Predictions for $log(y)$
- Standard Errors of Fitted Values

# Fitted Values

(Review) Already in this course we've discussed making predictions from our models and our data. Early in the semester we called them 'fitted values' and later we used them in thinking about interpreting interaction models.

If we estimated a model such that:

$$\hat{y} = 11 + 2x_1 + 4x_2 + 0.25x_3$$

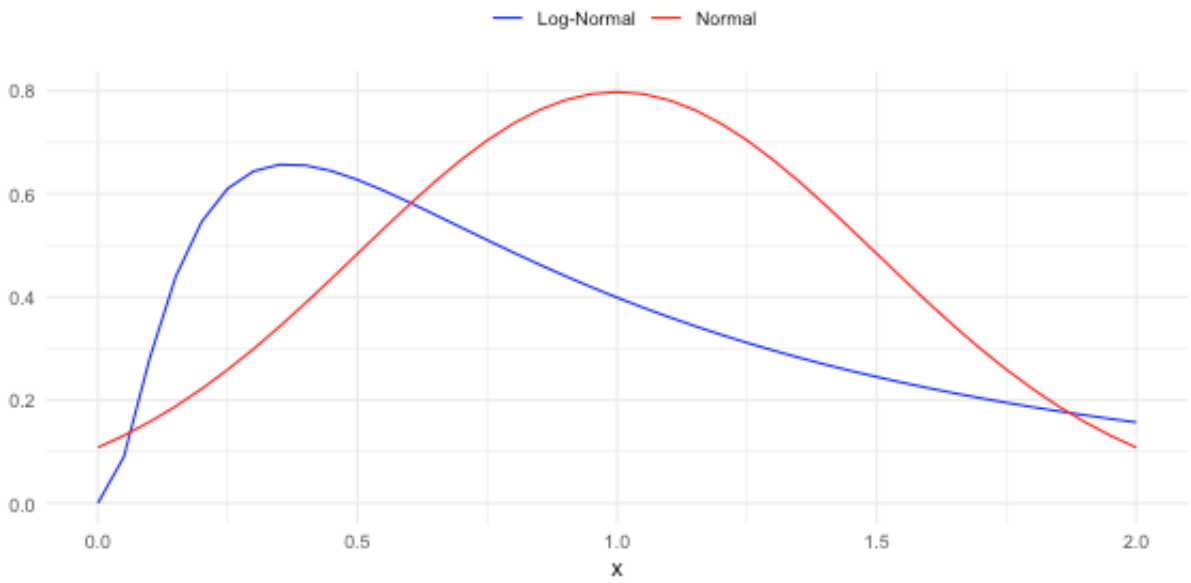
we could find the following *fitted values* (also known as *predictions*) from our data.

| x1 | x2 | x3 | y_i | y_pred |
|---|---|---|---|---|
| 1 | 2 | 1 | 12 | 23.25 |
| 5 | 3 | 2 | 30 | 36.50 |
| 3 | 0 | 3 | 9 | 17.75 |

# Adapting Predictions for $log(y)$

$$log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

(Review) Due to our errors being distributed according to a log-normal distribution, predictions from a log-level or log-log model are going to under-estimate the expected value of $y$ when we convert our $\hat{log(y)}$ to $y$.



# Adapting Predictions for $log(y)$

In other words, the expected method

$$E(y|X) = e^{\hat{log(y)}}$$

would create underestimates.

To fix this, we add a correction factor such that:

$$E(y|X) = \hat{\alpha}_0 e^{\hat{log(y)}}$$

where we estimate $\hat{\alpha}$ by looking at the average exponentiated residual or $\breve{\alpha}$ by standardizing the exponentiated values by the raw values of $y$.

# Standard Error of the Prediction

(Review) Estimating a prediction is easy (plug it in the equation!). Estimating the standard error is a bit more complicated.

If we have an estimated equation such that:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

then the parameter we aim to estimate is:

$$\theta_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \ldots + \beta_k c_k$$

where $c$ represents particular values (or constants) for each of $k$ independent variables.

Finally, our estimator of $\theta_0$ becomes:

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \ldots + \hat{\beta}_k c_k$$

# Standard Error of the Prediction

To obtain a standard error for $\hat{\theta}_0$ , we need to consider the linear combination of our OLS estimators (because the value is dependent on all values of $\hat{\beta}_j$, unless $c_j$ = 0).

To find the standard error associated with the *expected value* of $y$, we can:

1. re-arrange our parameter equation so that

$$\theta_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \ldots + \beta_k c_k$$

$$\beta_0 = \theta_0 - \beta_1 c_1 - \beta_2 c_2 - \ldots - \beta_k c_k$$

2. substitute our rearranged parameter equation into our regression formula so that

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

$$y = \theta_0 + \beta_1 (x_1 - c_1) + \beta_2 (x_2 - c_2) + \ldots + \beta_k (x_k - c_k)$$

... which is the equivalent to estimating a regression for $y$ on $x_1 - c_1$, $x_2 - c_2$ and using the standard error from the intercept term!

# Prediction Examples

| | **Clinton Share** |
|---|---|
| Intercept | 54.196*** |
| | (2.014) |
| Median Age | -0.106* |
| | (0.054) |
| Per Capita Income | -0.001*** |
| | (0.000) |
| $R^2$ | 0.052 |
| Adj. $R^2$ | 0.051 |
| Num. obs. | 2704 |

***p < 0.001; **p < 0.01; *p < 0.05

Statistical models

# Prediction Examples

| | Clinton Share | Clinton Share |
|---|---|---|
| Intercept | 54.196*** | 43.807*** |
| | (2.014) | (0.413) |
| Median Age | -0.106* | |
| | (0.054) | |
| Per Capita Income | -0.001*** | |
| | (0.000) | |
| Median Age - 35 | | -0.106* |
| | | (0.054) |
| Per Capita Income - 10000 | | -0.001*** |
| | | (0.000) |
| $R^2$ | 0.052 | 0.052 |
| Adj. $R^2$ | 0.051 | 0.051 |
| Num. obs. | 2704 | 2704 |

***p < 0.001; **p < 0.01; *p < 0.05

Statistical models

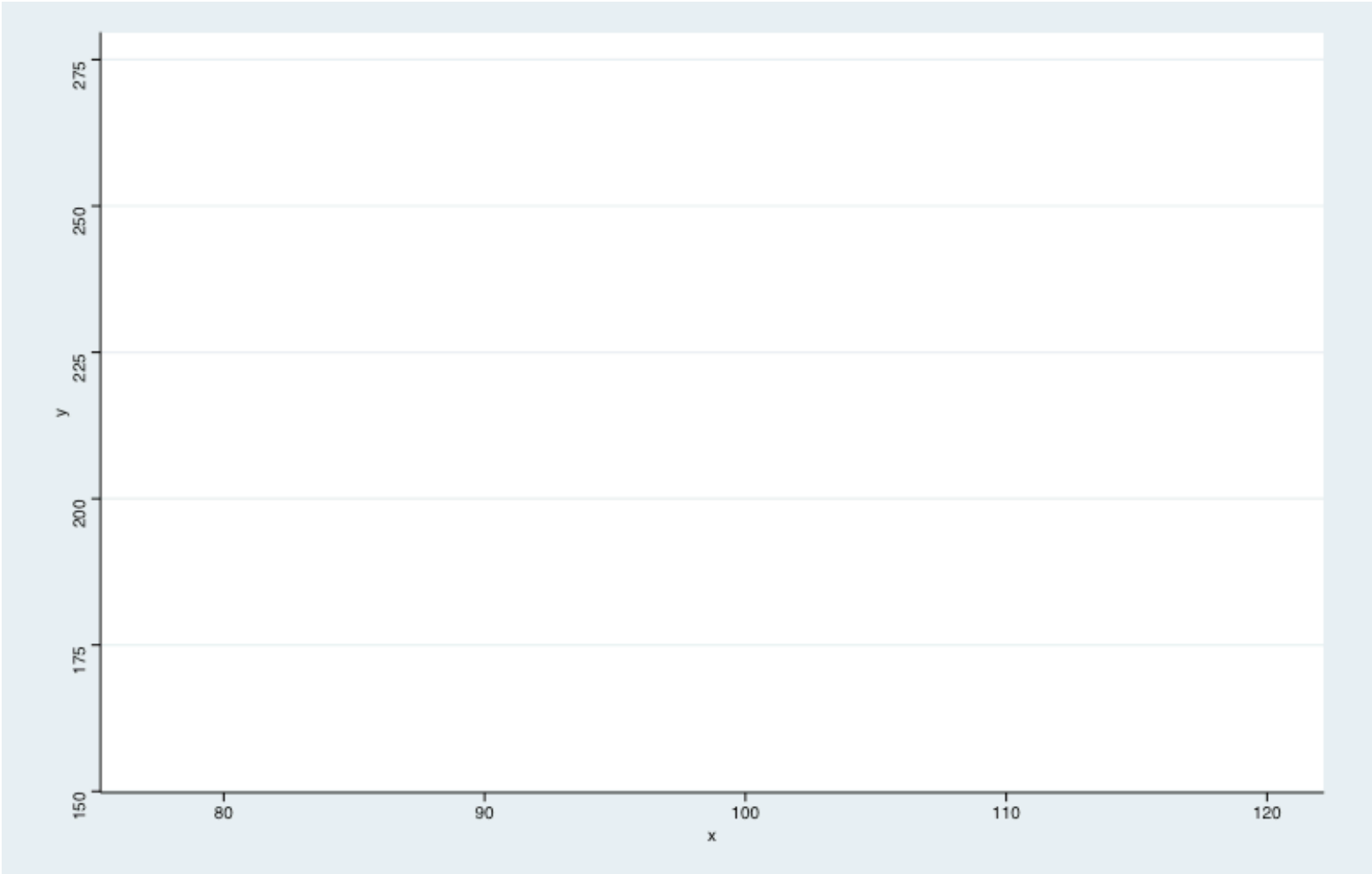
... which provides the prediction (and significance) for when median age = 35 and per capita income = $10,000 (i.e. the predictions of the expected value of our dependent variable given these covariates!)

# Prediction Interval

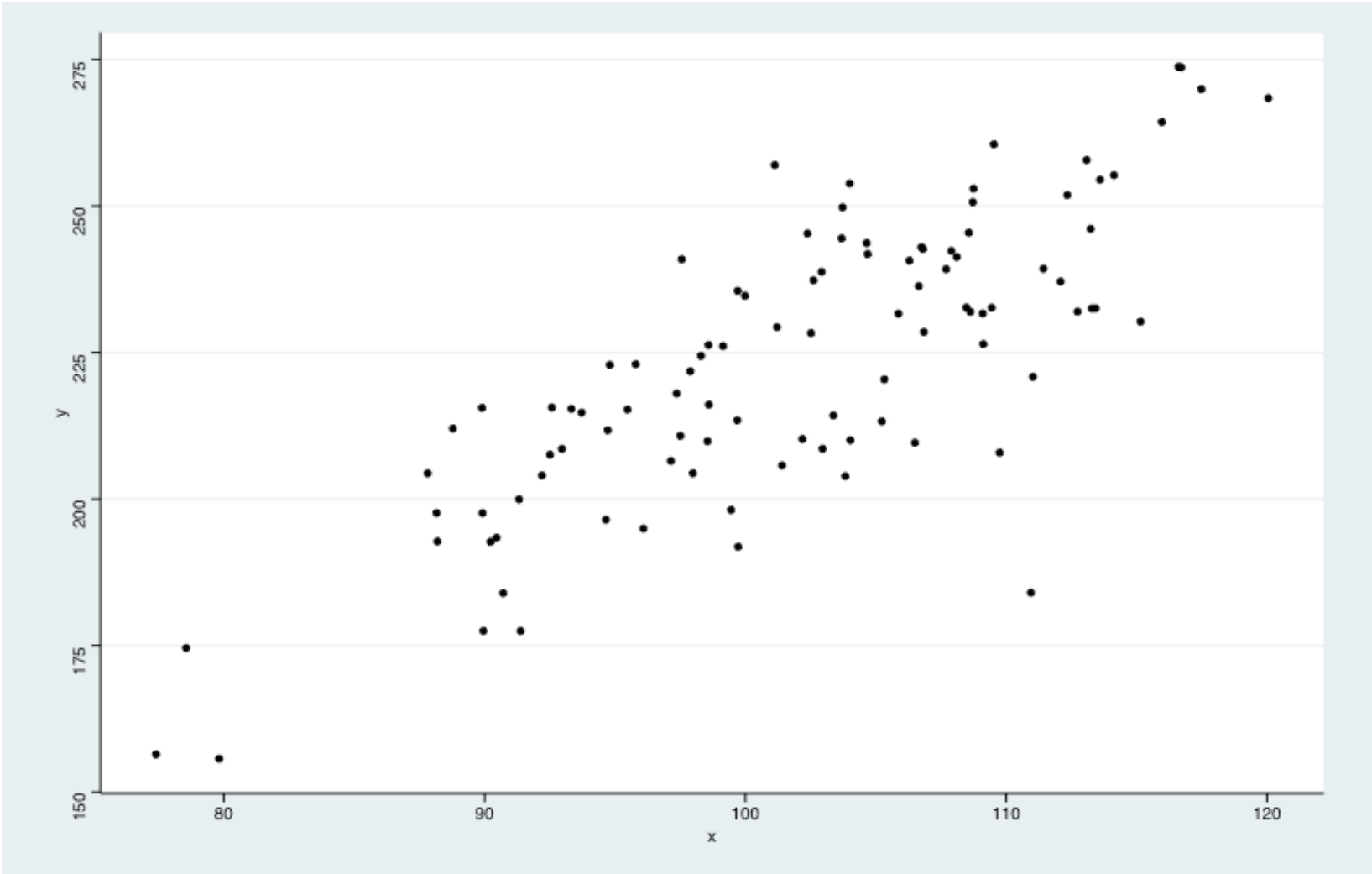
The previous slides are useful by producing the standard error of $\hat{y}$.

However, when wrestling with predictions, we might be interested not in the standard error of the *expectation* but rather the standard error of our *predictions*.

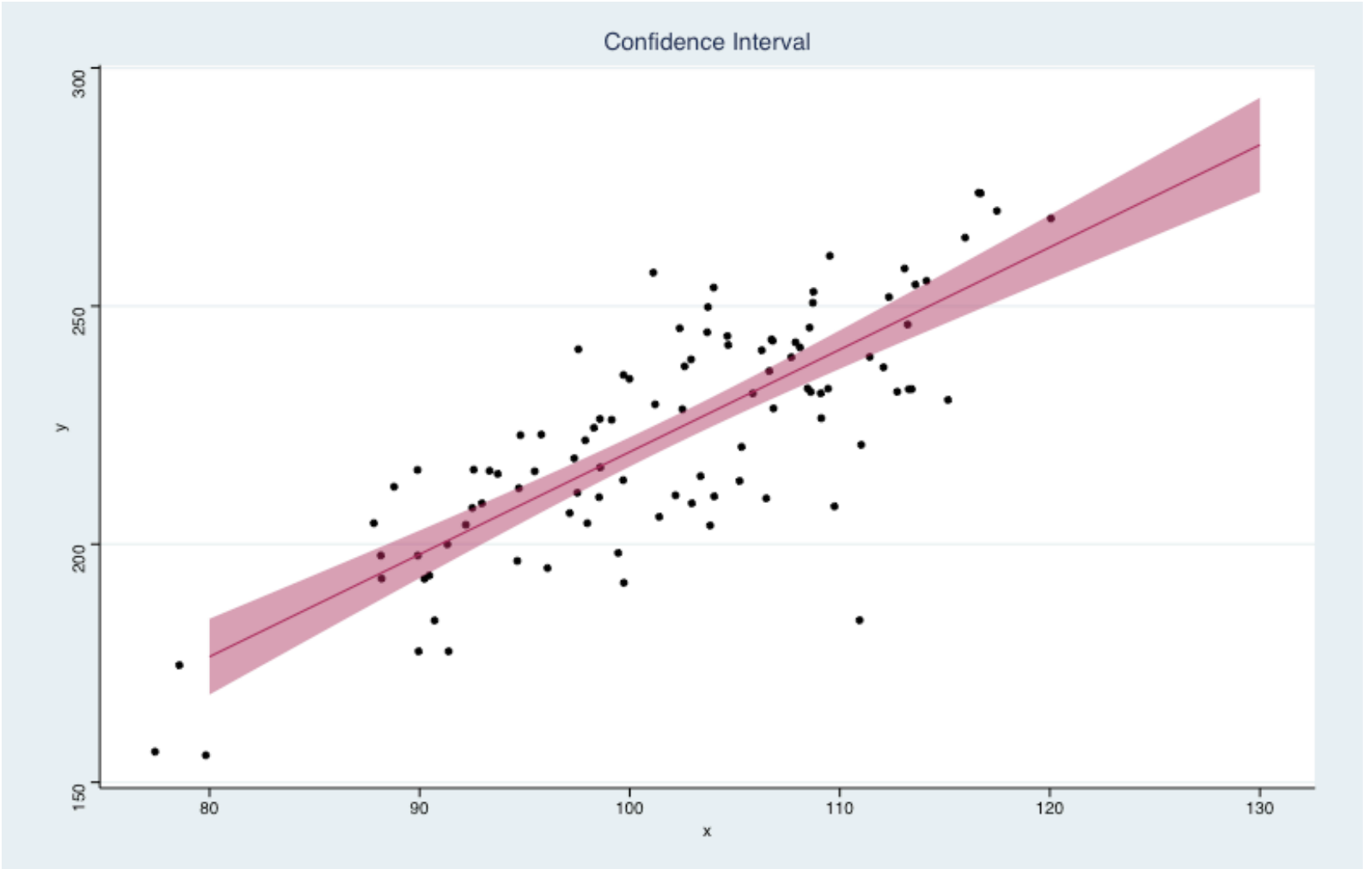
# Prediction Interval



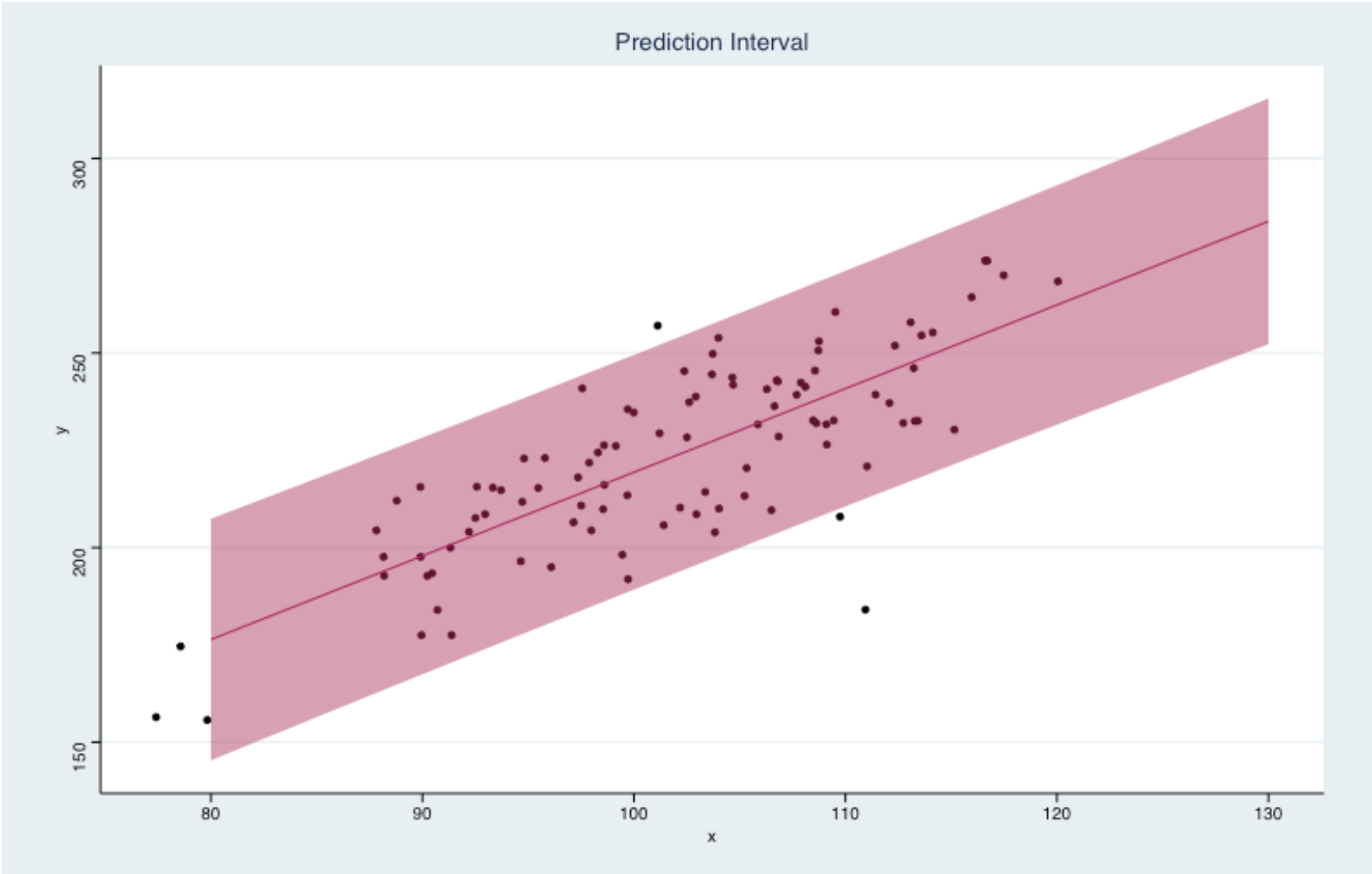
# Prediction Interval



# Prediction Interval



# Prediction Interval



# Prediction Interval

Generally speaking, the **confidence interval** refers to the range of *expected* values. A **prediction interval** refers to the fuller uncertainty associated with a single value.

Assume we have two intervals, both at 95%:

- the confidence interval band reveals the 95% confidence of the interval holding the expected value (think about sampling error)
- the prediction interval reveals the 95% confidence of the interval containing various predictions (think sampling error + population error)

In other words, the prediction interval accounts for additional variance in unobserved error

# Calculating the Prediction Interval

As mentioned on the previous slide, our prediction interval is the sum of sampling error and population error, or:

$$Var(\hat{e^0}) = Var(\hat{y^0}) + Var(u^0) = Var(\hat{y^0}) + \sigma^2$$

Expressed slightly differently, we tend to estimate a standard error of the prediction error, or:

$$se(\hat{e}^0) = \sqrt{se(\hat{y^0})^2 + \hat{\sigma^2}}$$

where:

- $se(\hat{y^0})$ is the estimated standard error of our prediction $\hat{y^0}$}
- $\hat{\sigma^2}$ is the variance of the residuals

which we often express as an interval estimated in the familiar way as:

$$\hat{y^0} \pm t * se(\hat{e}^0)$$

# Prediction Interval Example

| | Clinton Share | Clinton Share |
|---|---|---|
| Intercept | 54.196*** | 43.807*** |
| | (2.014) | (0.413) |
| Median Age | -0.106* | |
| | (0.054) | |
| Per Capita Income | -0.001*** | |
| | (0.000) | |
| Median Age - 35 | | -0.106* |
| | | (0.054) |
| Per Capita Income - 10000 | | -0.001*** |
| | | (0.000) |
| $R^2$ | 0.052 | 0.052 |
| Adj. $R^2$ | 0.051 | 0.051 |
| Num. obs. | 2704 | 2704 |
| RMSE | 9.918 | 9.918 |

***p < 0.001; **p < 0.01; *p < 0.05

Statistical models

# Prediction Interval Example

If we want to find the prediction interval where median age = 35 and per capita income is 10000, we can solve using:

$$se(\hat{e}^0) = \sqrt{se(\hat{y^0})^2 + \hat{\sigma^2}}$$

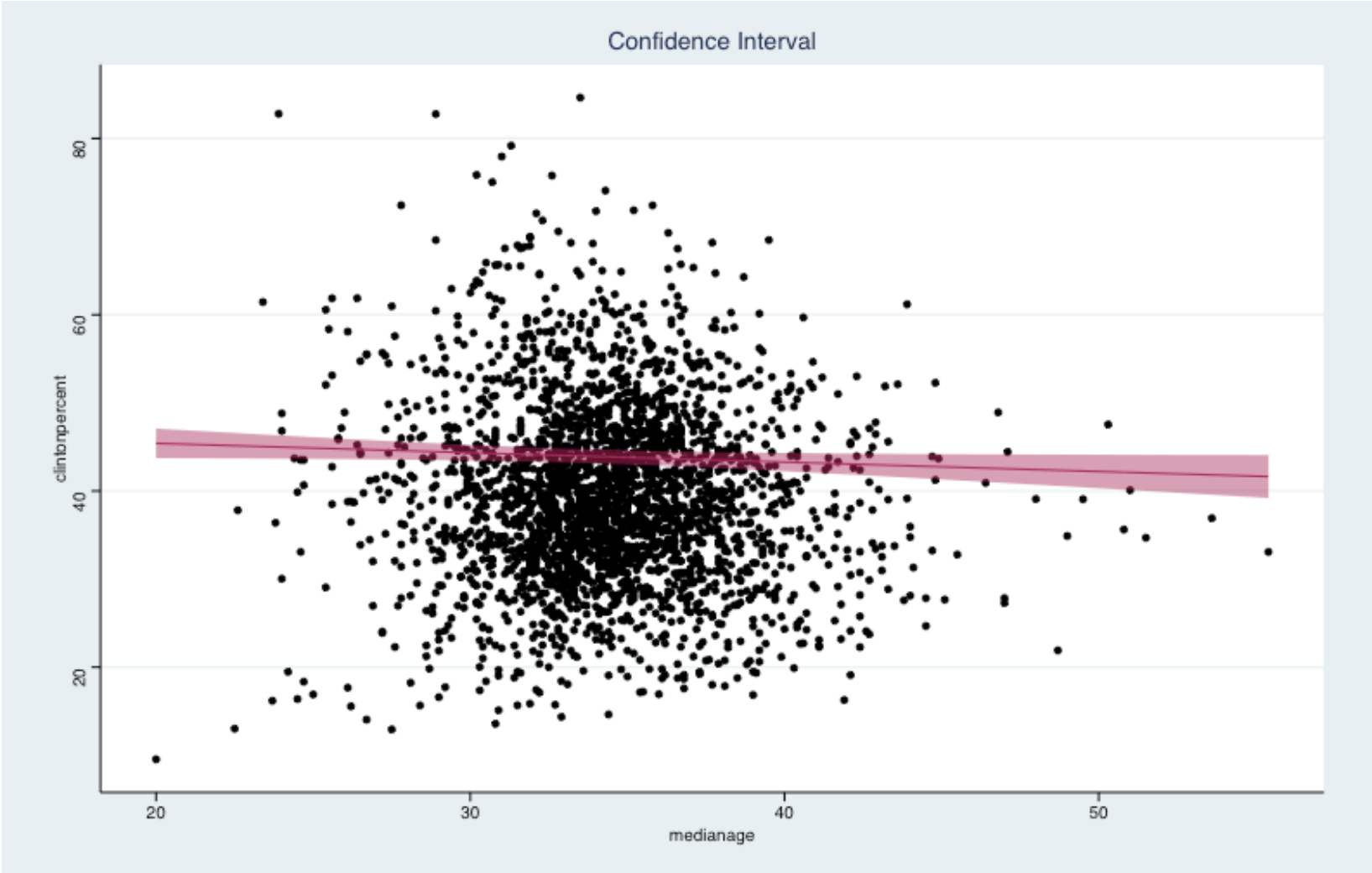
$$se(\hat{y^0}) = 0.413$$

$$\hat{\sigma} = 9.918$$

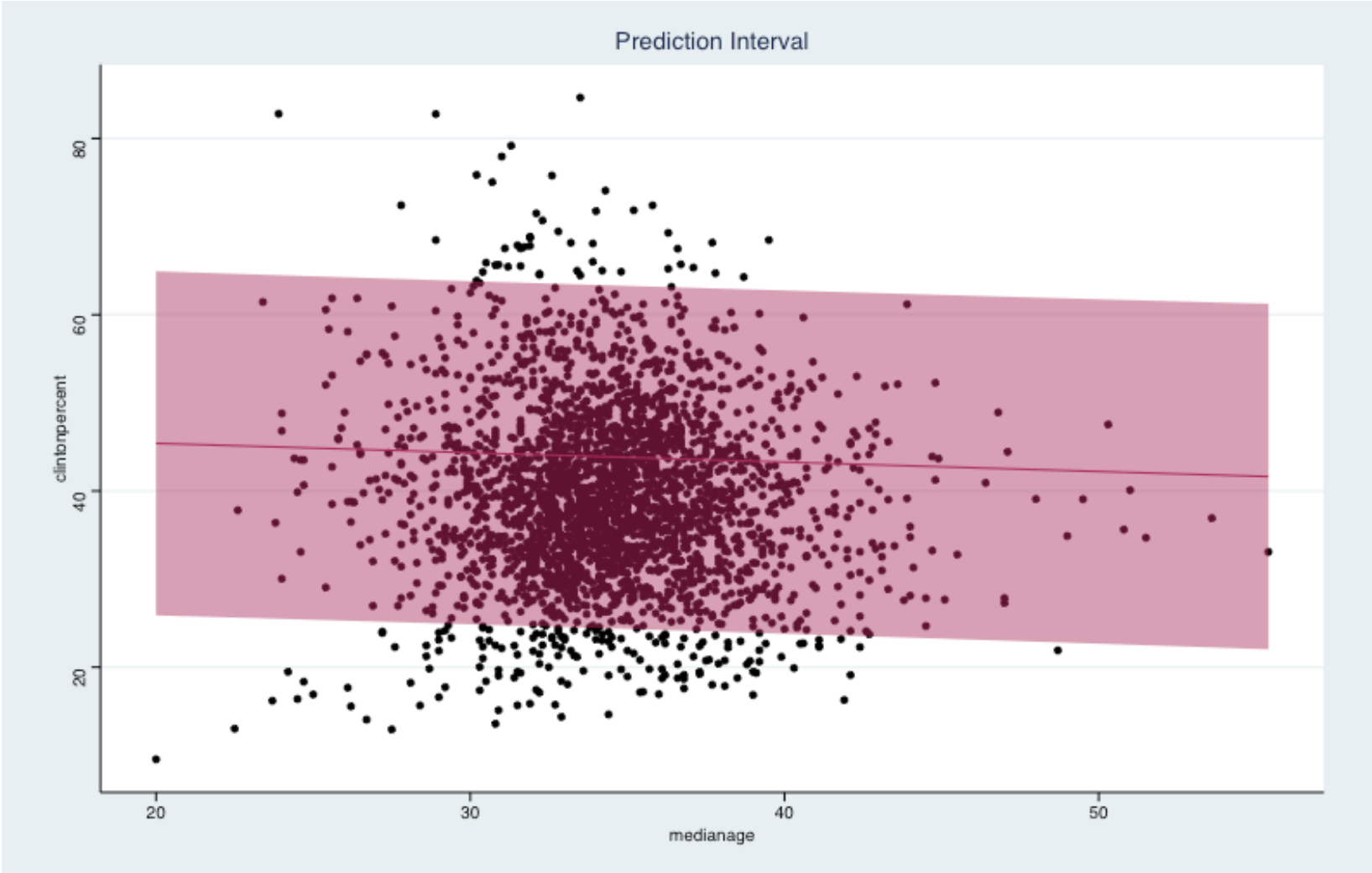
$$se(\hat{e}^0) = \sqrt{0.413^2 + 9.918^2}$$

$$se(\hat{e}^0) = 9.926$$

# Prediction Interval Example



# Prediction Interval Example



# Using Predictions

To take a step back, there is often a distinction between **explanatory modeling** and **predictive modeling** (Schmueli 2020).

- Explanatory modeling - the use of statistical models for testing causal explanations, where we can make inferences to test hypotheses
- Predictive modeling - applying a statistical model or algorithm to data for the purpose of predicting new (or future) observations

A useful framework for distinguishing between the two is identifying the focus of the following axes:

- *causation-association* - explanatory seeks the causal; predictive identifies the association
- *theory-data* - theory often dictates explanatory models; data often drives predictive models
- *retrospective-prospective* - explanatory is backward looking; predictive is forward looking
- *bias-variance* - explanatory seeks to minimize bias; predictive to minimize variance

# Using Predictions

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Using the same model, we might focus on different elements:

- explanatory would likely focus on the estimates for $\beta$
- predictive would likely focus on how well we can estimate the fitted values

# Working with Out of Sample Data

So far in class we've mostly just spoken about our "data" and we've used that data to estimate various models for explanatory purposes. But as we pivot from explanatory to predictive we also need to think of our data differently.

One way to think of our data is to reference "in-sample" vs. "out-of-sample" data. Generally this means refers to whether or not we collected and used a particular datum to estimate our model.

# Working with Out of Sample Data: Example 1

Imagine we estimate the following equation:

$$ViolentC = \beta_0 + \beta_1 PoliceFunding + \beta_2 RegB + \beta_2 RegC + \beta_2 RegD + \beta_2 RegE$$

| | OLS |
|---|---|
| Intercept | -392.982 |
| | (240.271) |
| Annual Police Funding | 16.883** |
| | (5.099) |
| Region B | 122.952 |
| | (199.184) |
| Region C | 400.832* |
| | (195.919) |
| Region D | 684.849** |
| | (195.454) |
| Region E | 649.676** |
| | (202.627) |
| $R^2$ | 0.482 |
| Adj. $R^2$ | 0.423 |

# Working with Out of Sample Data: Example 1

Out of sample data might help us answer...

*What if State 41 in Region E increased its annual police funding by five points?*

Current value: 3545

Current fitted value: 1708

New fitted value: 1793

Confidence interval: [1284, 2301]

Prediction interval: [778, 2807]

*What if a new state were added to region B and had the maximum annual police funding?*

Current value: NA

Current fitted value: NA

New fitted value: 1181

# Working with Out of Sample Data: Example 2

To explore a second example, now imagine using our admissions data we have estimated the model:

$$P(Admitted = 1) = \Phi(\beta_0 + \beta_1 GRE + \beta_2 CGPA + \beta_3 Research)$$

| | Probit |
|---|---|
| Intercept | -33.696*** |
| | (3.588) |
| GRE Score | 0.055*** |
| | (0.013) |
| CGPA | 1.914*** |
| | (0.264) |
| Research | 0.466** |
| | (0.171) |
| AIC | 319.544 |
| BIC | 336.403 |
| Log Likelihood | -155.772 |
| Deviance | 311.544 |
| Num. obs. | 500 |

***p < 0.001; **p < 0.01; *p < 0.05

# Working with Out of Sample Data: Example 2

*What does the model predict for a student with a GRE score of 330, a GPA of 8, and no research experience?*

Probability: 44%

# Fitting the Model

We may be tempted when building a model to have the *best* fit possible (think: highest $R^2$ or lowest $RMSE$, however that might not always be the smartest decision.

In part, because we might accidentally come up with the perfect model for our *sample* but not for our *population*.

# Fitting the Model

Let's take an example where we have 100 cases in our population. We *know* the relationship is:

$$y = 1 + .4x + x^2 + u$$

but let's say we sample 40 cases and try to estimate the full relationship. We might consider the following models:

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| Intercept | -15.287* | -4.625 | -2.101 | -2.627 |
| | (6.414) | (4.356) | (4.524) | (5.662) |
| X | 11.024*** | 1.204 | 3.314 | 3.339 |
| | (0.988) | (1.464) | (1.918) | (1.951) |
| X^2 | | 0.994*** | 0.286 | 0.390 |
| | | (0.134) | (0.449) | (0.800) |
| X^3 | | | 0.045 | 0.025 |
| | | | (0.027) | (0.130) |
| X^4 | | | | 0.001 |
| | | | | (0.006) |
| $R^2$ | 0.766 | 0.906 | 0.913 | 0.913 |
| Adj. $R^2$ | 0.760 | 0.901 | 0.906 | 0.903 |
| Num. obs. | 40 | 40 | 40 | 40 |

# Fitting the Model

We might assume that model four is the best model for making predictions considering the high $R^2$ and adjusted $R^2$.

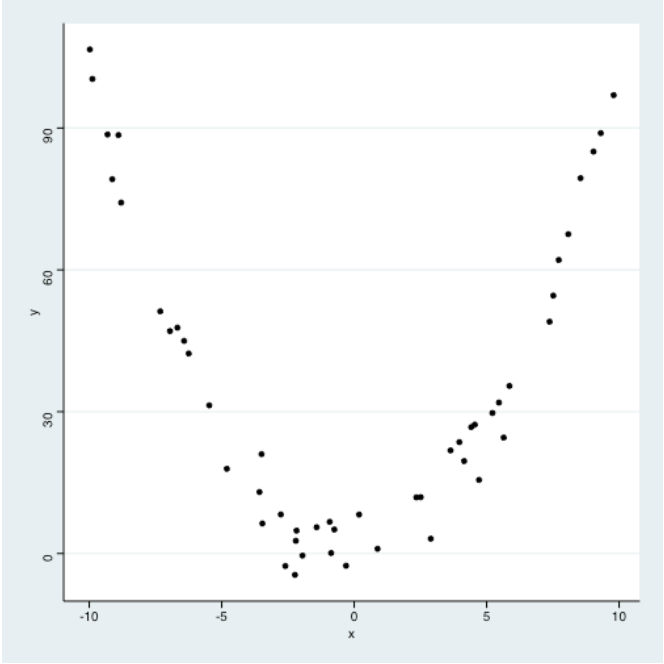
But let's consider which model performs best on a *different* random sample from our population. Here, we'll sample 20 new observations and consider the RMSE. We can compare the RMSE from this out-of-sample group to the in-sample.

| | RMSE Mod 1 | RMSE Mod 2 | RMSE Mod 3 | RMSE Mod 4 |
|---|---|---|---|---|
| Original 40 | 15.10489 | 11.4661 | 11.45303 | 11.08218 |
| New 20 | 35.18467 | 13.4988 | 12.94296 | 57.99199 |

# Overfitting and Underfitting

This example introduces us to:

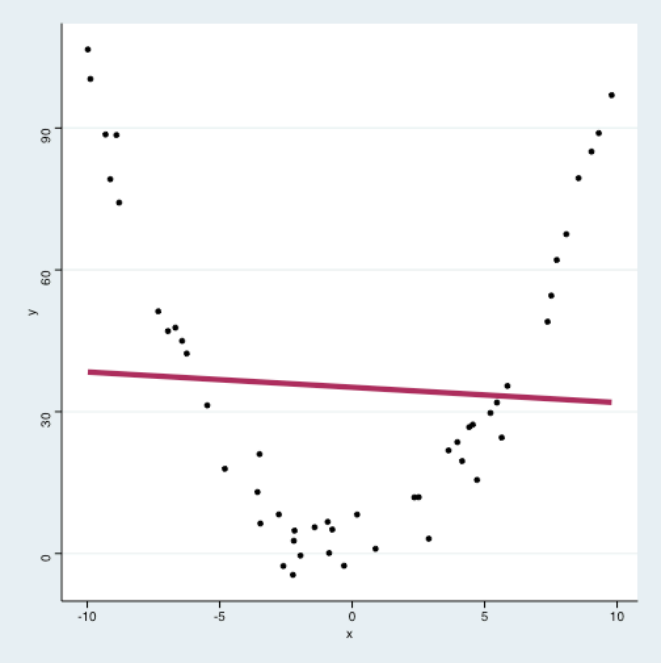
- underfitting: not great fit to the training data, not great fit to other data (high bias)
- overfitting: great fit to our training data, not as great generalization to other data (such as the testing set) (high variance)



# Overfitting and Underfitting

This example introduces us to:

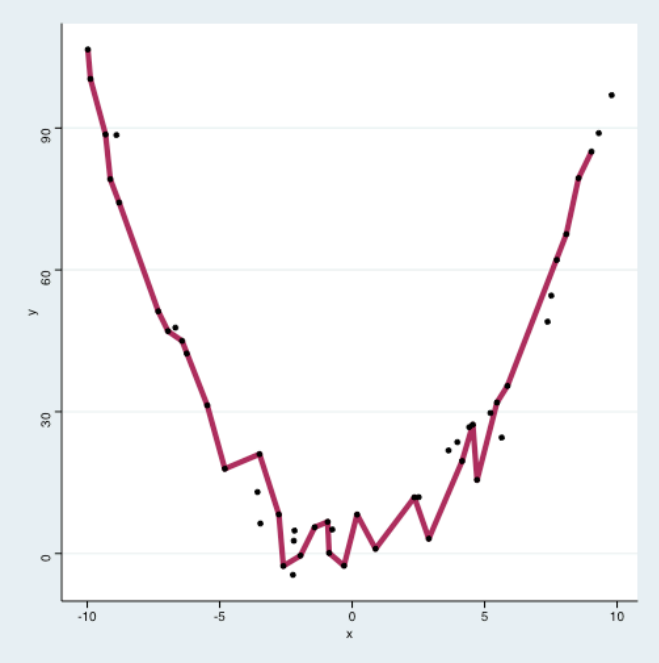
- **underfitting**: not great fit to the training data, not great fit to other data (high bias)
- overfitting: great fit to our training data, not as great generalization to other data (such as the testing set) (high variance)



# Overfitting and Underfitting

This example introduces us to:

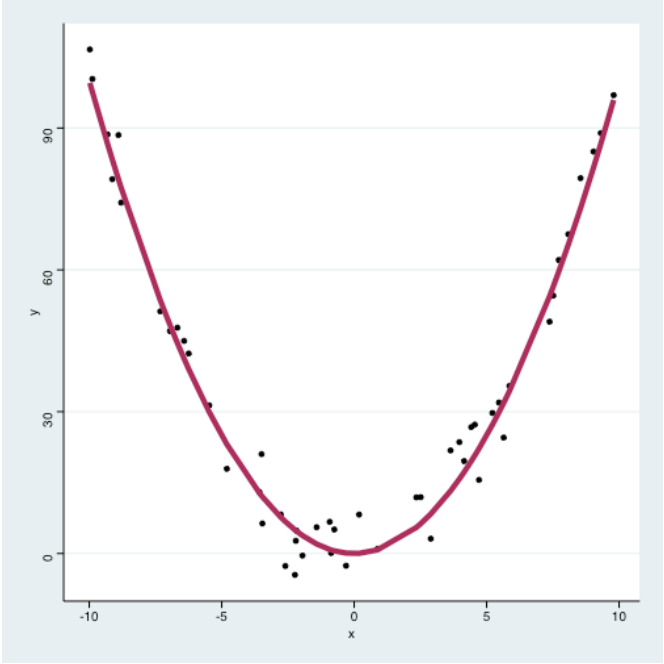
- underfitting: not great fit to the training data, not great fit to other data (high bias)
- **overfitting**: great fit to our training data, not as great generalization to other data (such as the testing set) (high variance)



# Overfitting and Underfitting

This example introduces us to:

- underfitting: not great fit to the training data, not great fit to other data (high bias)
- overfitting: great fit to our training data, not as great generalization to other data (such as the testing set) (high variance)



# Prediction Workflow

The workflow for creating predictive models is *slightly* different than how we've discussed in the past -- in part because we are now quite sensitive to the accuracy of out-of-sample predictions.

To prevent overfitting, we will often follow:

- Randomly dividing our data into training data and testing data
- Estimate a model using our training data (and *only* our training data)
- Use this model (estimated by the training data) to make predictions on the testing data
- Assess the accuracy of our model by comparing the predictions with the actual observed values. And this workflow works for OLS, logit, probit, and many others.

- We're *willingly* using less data to estimate our model (gasp!)
- We're no longer just using in-sample or hypothetical data points to render predictions.

# Prediction Workflow Example

Let's say we have a data set of the GSS and we're interested in developing a simple model to predict whether someone voted (in the 2008 election).

Show 6 entries Search:

| | year | id | wrkstat | wrkslf | wrkgvt | marital | sibs | childs |
|---|---|---|---|---|---|---|---|---|
| 1 | 2012 | 1 | WORKING PART TIME | SOMEONE ELSE | PRIVATE | Never married | 1 | 0 |
| 2 | 2012 | 2 | WORKING PART TIME | SOMEONE ELSE | PRIVATE | Never married | 2 | 0 |
| 3 | 2012 | 3 | WORKING FULL TIME | SOMEONE ELSE | PRIVATE | Married | 1 | 2 |
| | | | | SOMEONE | | | | |

# Prediction Workflow Example

**Step 1:** We're going to randomly assign 80% of our data to be the "training set" and the remaining 20% of our data as the "test set".

This leaves our training data with 1580 cases and our testing data with 394 cases.

**Step 2:** We can estimate a model using our training data (and only our training data)

$$P(Vote = 1) = \Lambda(\beta_0 + \beta_1 Age + \beta_2 Educ + \beta_3 Sex + \beta_4 Race + \beta_5 Party + \beta_6 Income)$$

**Step 3:** We will use that model estimated in step 2 and make predictions to our test data

# Prediction Workflow Example

**Step 4:** Assess the accuracy of the model!

Since we're using a simple logit model, let's use a different metric (more on this later).

$$Accuracy = \frac{CorrectPredictions}{AllPredictions}$$

$$Accuracy = \frac{CorrectPredictions}{AllPredictions}$$

$$Accuracy = .745$$

# K-fold Cross Validation

**K-fold cross-validation** builds on the validation method and divides the dataset into $k$ number of subsets and computes our performance metric K number of times using each subset as a testing set.

For example, imagine we have a model and dataset that we want to use k-fold cross-validation where $k = 5$.

1. Randomly split the data into 5 subsets (A, B, C, D, and E).
2. For the first iteration, hold out subset A as the test set and use the remaining subsets (B, C, D, and E) as the training set.
3. Test the model that you developed on the remaining subsets on the test set and store the diagnostic we care about (e.g. prediction accuracy).
4. Repeat steps 2 and 3 for each subset.
5. Average the stored diagnostics and report it as the "cross validation performance metric."

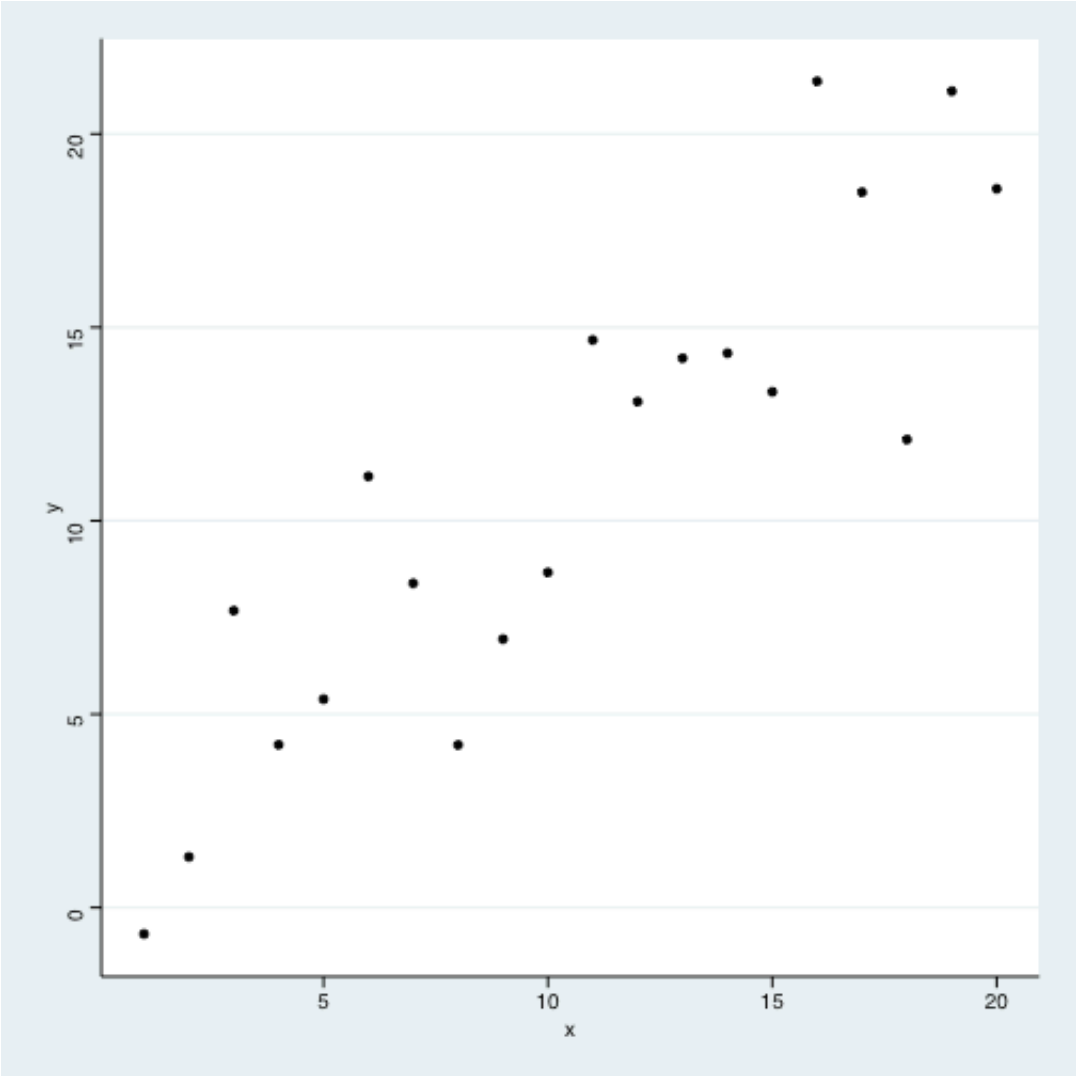
*Why is this better than the validation set method?:*

More robust due to repeated randomization.

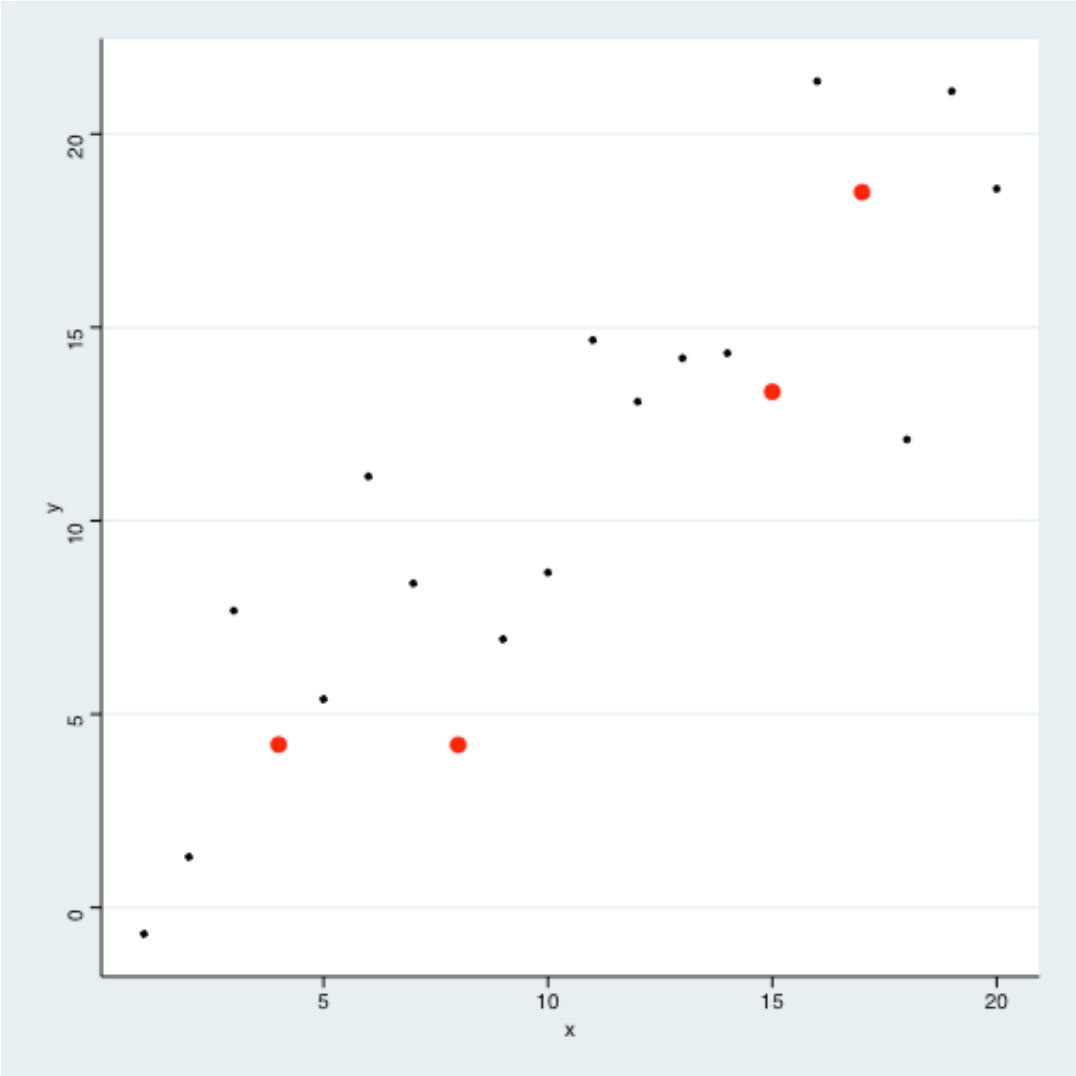
*How do we decide $k$?*

Typically, you will see k as 5 or 10 as those figures have been seen to not sacrifice bias at the expense of variance (or vice versa). i.e. Lower k values may be more biased but

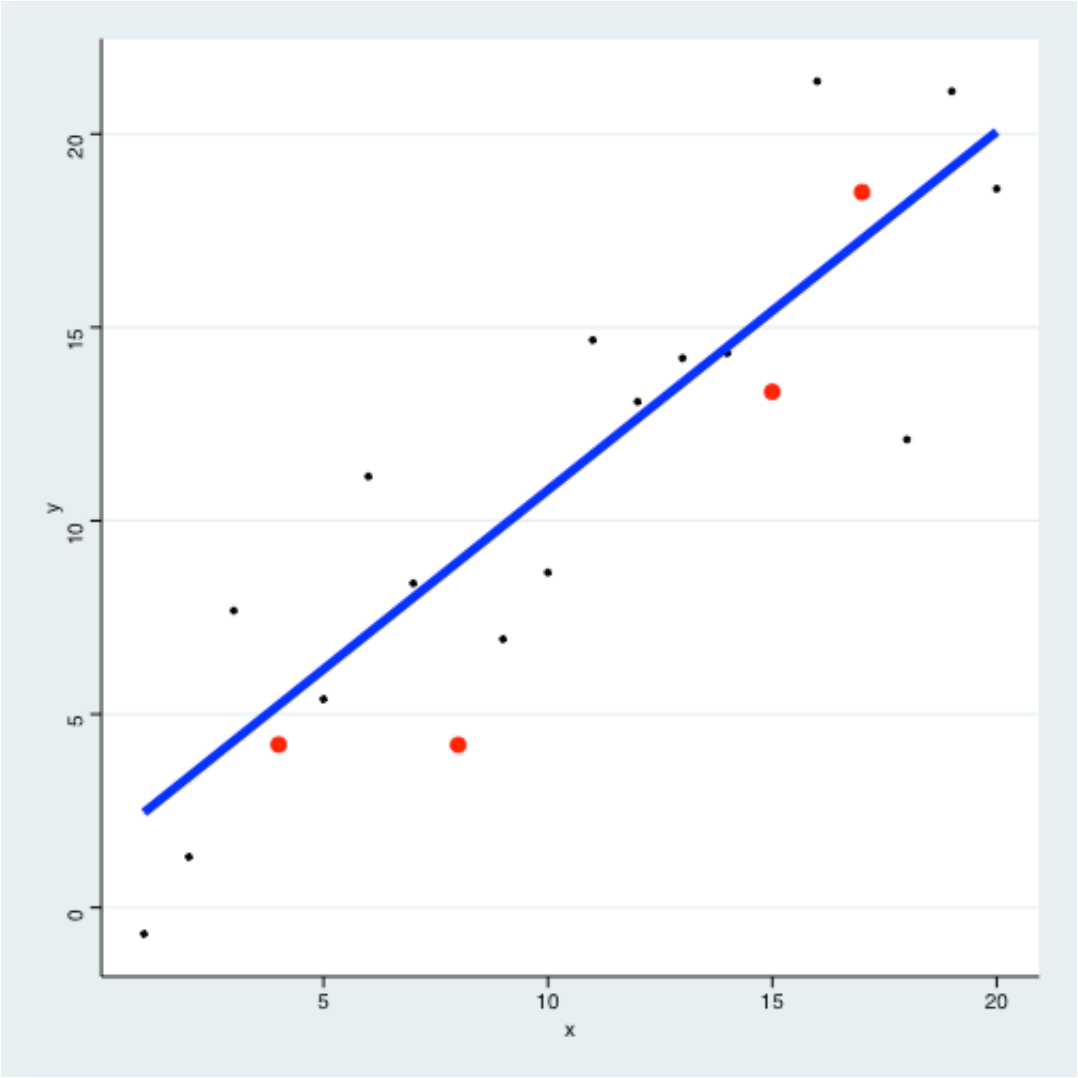
# K-fold Cross Validation (K = 5)



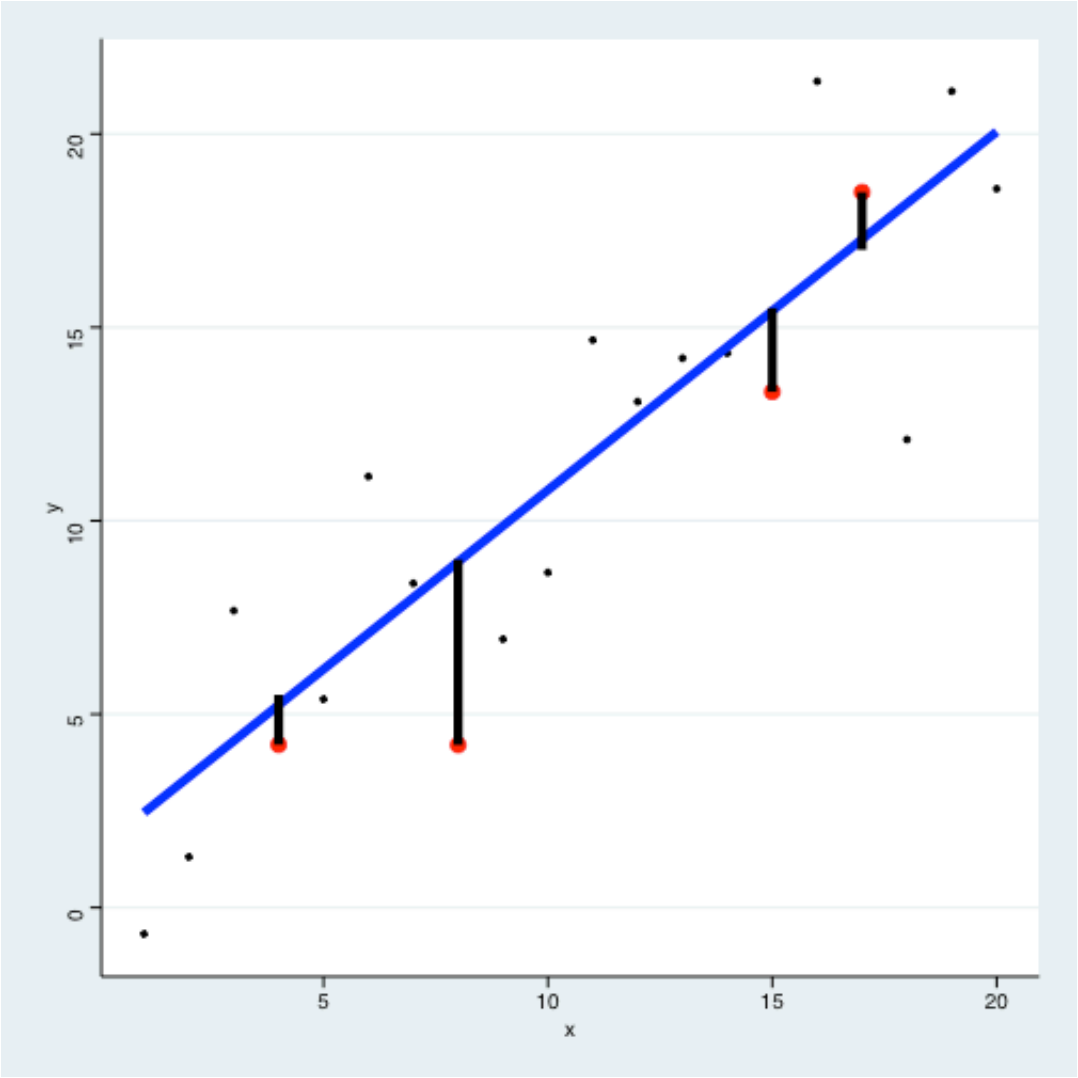
# K-fold Cross Validation (K = 5)



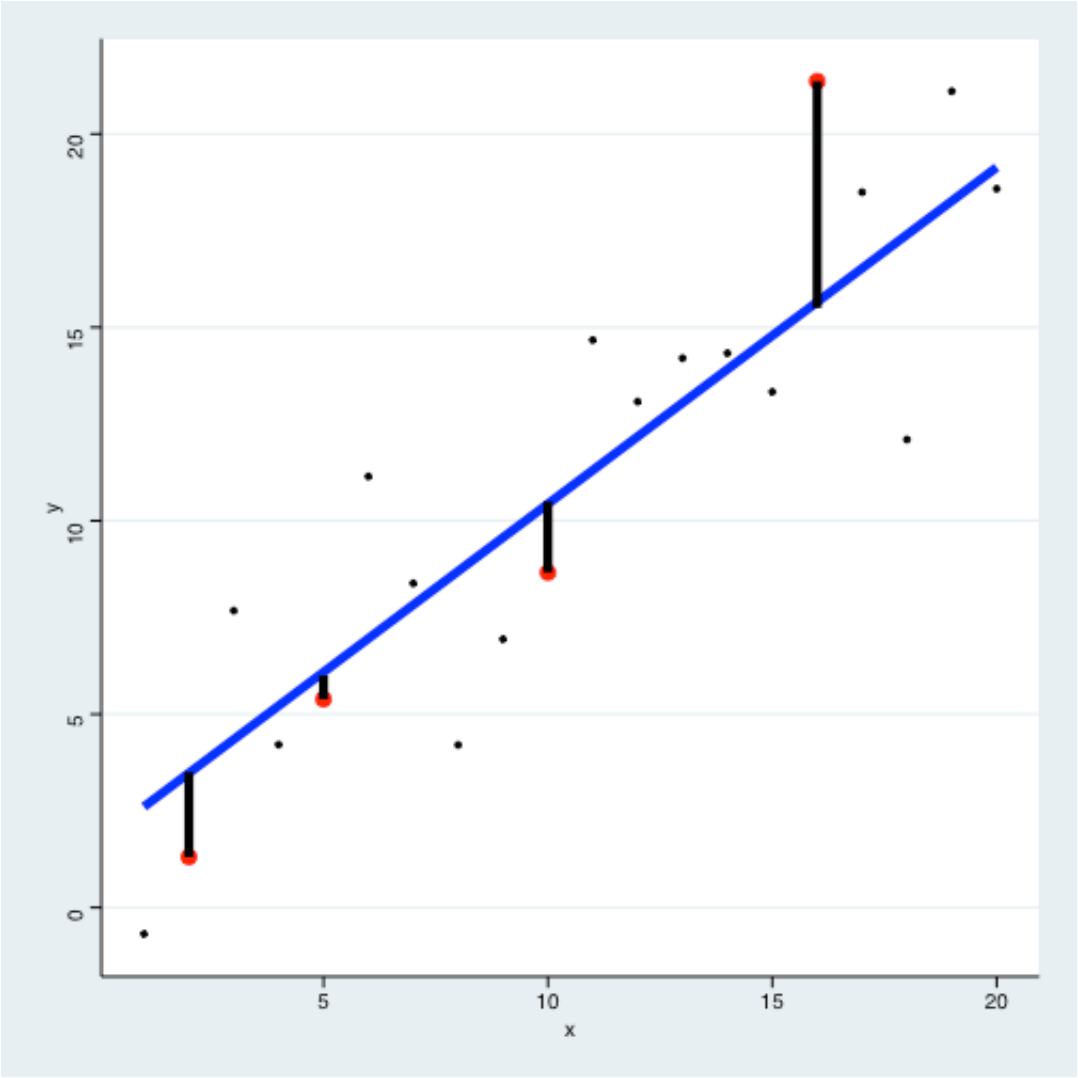
# K-fold Cross Validation (K = 5)



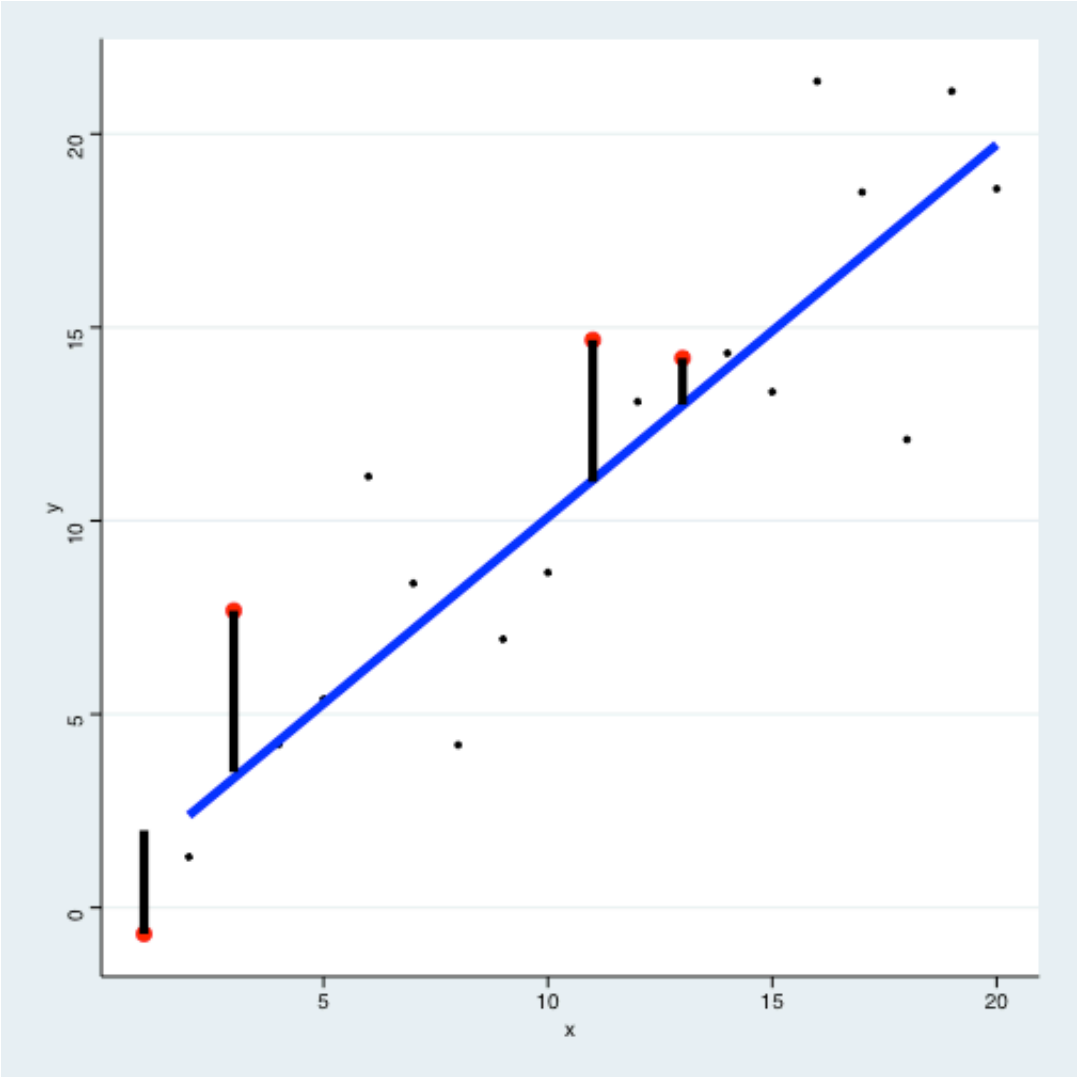
# K-fold Cross Validation (K = 5)



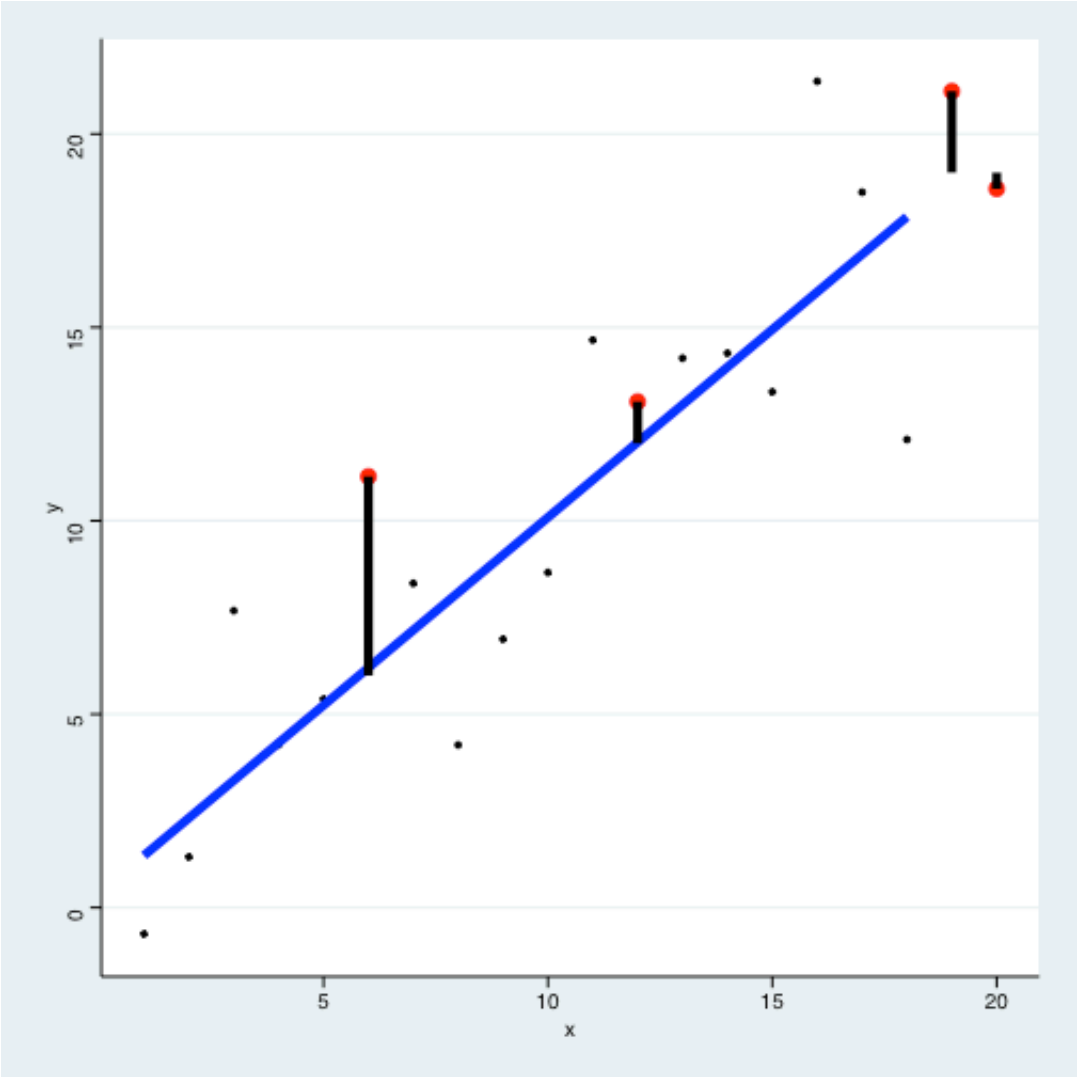
# K-fold Cross Validation (K = 5)



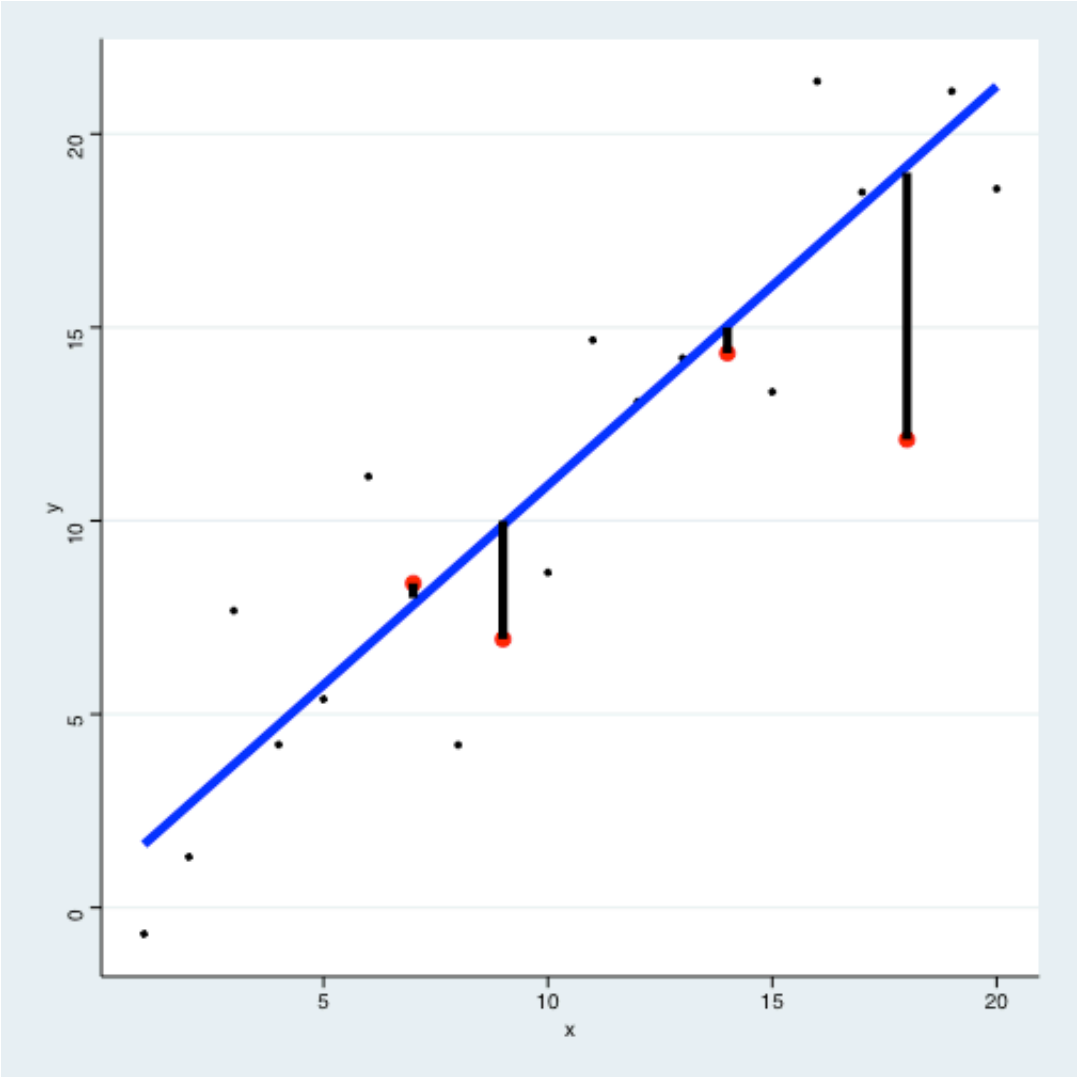
# K-fold Cross Validation (K = 5)



# K-fold Cross Validation (K = 5)



# K-fold Cross Validation (K = 5)



# Model Metrics

With the models that we've introduced in class (OLS, logit, probit), there are a number of different model metrics that we can use to assess model performance (think: goodness of fit). We're only going to scratch the surface here (and you don't need to memorize these at all) but...

**OLS** *(Continuous DV)*

$R^2$ or $Adj. R^2$

$RMSE$ (or just $MSE$)

$MAE$ or Mean Absolute Error, $= \frac{|y_i - y_p|}{n}$

$MAPE$ or Mean Absolute Percentage Error, $= \frac{100\%}{n} \sum \frac{|y_i - y_p|}{y_i}$

# Model Metrics

## **Logit/Probit** *(Binary DV)*

$PseudoR^2$

$Accuracy$ - Correct Predictions

$Specificty$ - Correctly Identified Negative Cases

$Sensitivity$ - Correctly Identified Positive Cases

$BrierScore$ - Evaluating the probability of the event, $= \frac{1}{N}\sum(p_i - y_i)^2$

Each has different benefits and disadvantages which we we'll pick up again during the Stata session.

# Applications

So why do we want to do this again?

Predictive models can get quite complicated but can begin with all the models that we've been estimating all semester long.

There are a lot of applications, many of which are explicitly working their way into the research sphere. Some examples include:

- developing microtargeting predictions for voters, political supporters, etc.
- working to predict outcomes of various policy implementations (think Covid vaccine policies)
- identifying opinion changes from smaller studies and extrapolating to broader populations

Though as models get more complicated, we have to be clear to think through the unintended consequences of our predictive models.

# Where We're Going Next

Missing data, and other data concerns

### Reading

- Wooldridge: 9-4, 9-5