When good models go bad: Model diagnostics, uncertainty, and assessment

POL 212: Quantitative Analysis I

Winter 2024

Diagnostics

Recommended (free!) text

Fox, John. 2020. *Regression Diagnostics*, 2nd ed. Thousand Oaks, CA: SAGE Publications.
https://doi.org/10.4135/9781071878651

Quick regression review

- What are we trying to do?
  1. Model some variable Y as a linear function of some variable(s) X.
  2. That is, we wish to regress Y onto X, estimating the population parameters α and β in the function: $Y_i = \alpha + \beta X_i + u_i$.
- We'd like these estimates to have a couple of properties:
  1. Minimize both bias (residuals) and sampling variance.
  2. Allow for statistical inference, modeling uncertainty in estimates.

The Gauss-Markov theorem

Turns out, the Ordinary Least Squares (OLS) estimator satisfies some (or all) of the points on our wish list *if* some assumptions are met.

- The first two (linearity and nonstochastic regressors) relate to bias.
- The next two (homoskedasticity and independence of errors) relate to variance.
- The above four assumptions are sufficient to prove OLS is BLUE (the Gauss-Markov theorem).
- We can add one more assumption (normality of errors) to complete the strong set and make statistical inference using the normal distribution.

What are these assumptions?

The weak set of Gauss-Markov assumptions (enough to prove OLS is BLUE):

1. **Linearity**: The expected (mean) value of the disturbance term is 0. Why important?
   - $E(Y_i) = E(Y|x_i) = E(\alpha + \beta_1 x_i + u_i)$
   - $\alpha + \beta_1 x_i + E(u_i)$
   - $\alpha + \beta_1 x_i$

What are these assumptions?

2. **Nonstochastic regressors, fixed X, or X independent of the error**: X values are independent of the error term (or X is **exogenous**). $cov(X_i, u_i) = 0$.
    - Can arise because of measurement error on X, omitted confounder(s), or simultaneous causation.
    - Example: economic conditions and civil conflict.
    - Instrumental variables: correlated with X, but not Y (e.g., rainfall).

What are these assumptions?

3. **Homoskedasticity**: constant error variance across values of X_i.
    - $Var(\varepsilon) = \sigma^2$
    - When violated? Situations where some values of X have greater uncertainty.
    - Robust standard errors.

What are these assumptions?

4. **Independence**: No autocorrelation between disturbances. $cov(u_i, u_j) = 0$ for $i \neq j$. Means OLS no longer efficient.
    - Influences our understanding of the error term.
    - Time series data.

OLS is unbiased if...

1. The functional form is correct, or the disturbance has a conditional mean of zero: $E(u_i|X_{1i}, X_{2i}, \ldots, X_{ki}) = 0, \forall i$.
2. The regressors are fixed, exogenous, or independent of the disturbance term:
   $cov(X_{1i}, u_i) = cov(X_{2i}, u_i) = \cdots = cov(X_{ki}, u_i) = 0$.

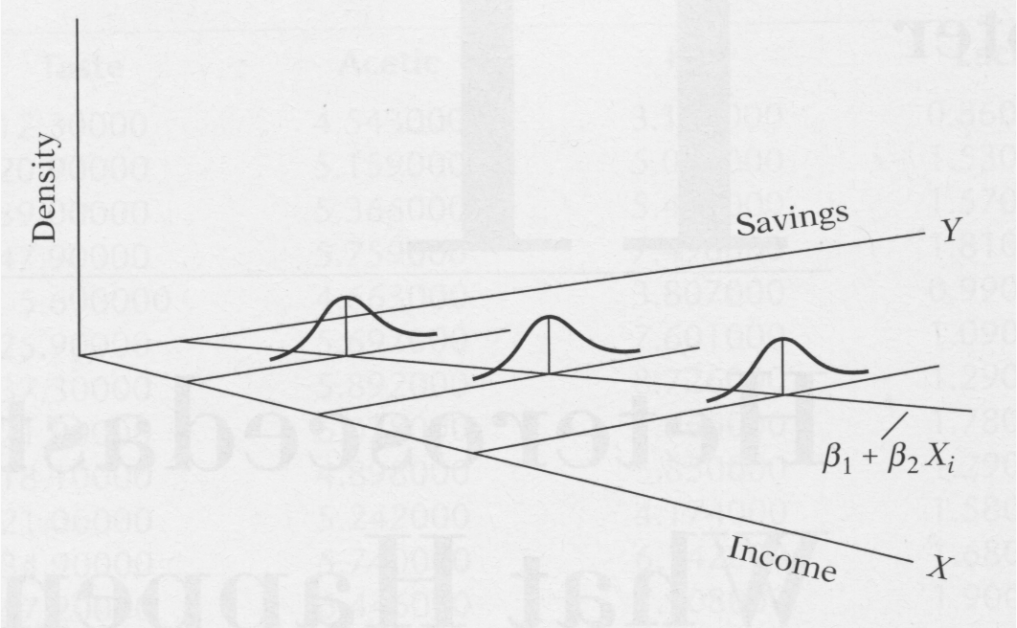
- Violating either of these two Gauss-Markov assumptions causes bias.

Possible Violations: Specification Errors

1. Omission of a relevant variable/wrong functional form
2. Errors of measurement
3. Incorrect specification of the error term
4. Reciprocal causation

Homoskedasticity and errors independent of each other

These don't deal with bias, but violations *do* mean that the estimates are no longer efficient (i.e., lowest variance of sampling distributions among all estimators in class).



And finally, a bonus (fifth) assumption...

- The assumption of normal disturbances completes the strong set: OLS is now the best unbiased estimator among all (not just linear) estimators (i.e., also the MLE estimate).
- The sampling variances of the estimators are now **normal**, allowing for easy and fun inference.
- Not possible if we're dealing with discrete dependent variables.
- **BUT**, no such restriction on the independent/predictor variables.

Assessing non-normality

- Non-normal errors: OLS no longer necessarily equivalent to MLE.
- Can assess by looking at (studentized) residuals: quantile-comparison plots and density of studentized residuals

Non-normal errors

```
ols1 <- lm(Life.Exp ~ Frost + Murder, data=states)
qqPlot(ols1)
plot(density(rstudent(ols1)))
```

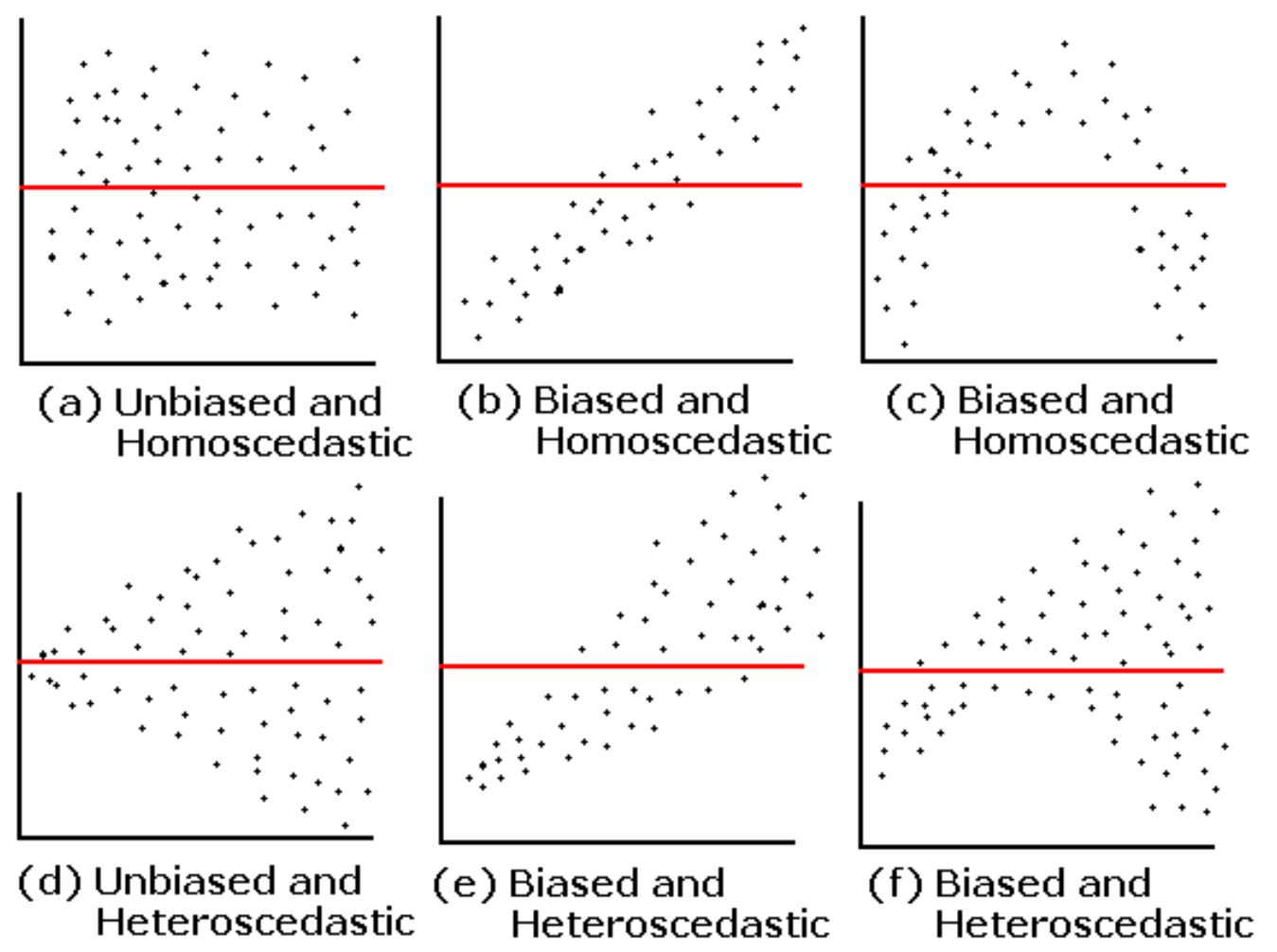
Assessing heteroskedasticity

- Non-constant error variance: bad standard errors. (Huber-White/robust SEs)
- Arises from incorrect specification (omitting an important effect on Y).
- Fitted vs. residual plots:

Residual plot

```
residualPlots(ols1, ~1, fitted=TRUE, tests=FALSE)
```

Residual plots



Assessing nonlinearity

- Nonlinearity: implies $E(\varepsilon) = 0$ everywhere.
- Might be caused by missing interaction or polynomial (or other transformation) terms.
- **Component-plus-residual plots** plot X_j against $\varepsilon + \hat{\beta}_j X_j$

Component-plus-residual plot

```
crPlots(ols1)
```

Multicollinearity and micronumerosity: what's the problem?

- We cannot estimate a regression model with OLS in a case of either:
  - Perfect multicollinearity, or one predictor being a perfect function of one or more other predictors.
  - Exact micronumerosity, or having fewer observations than parameters to be estimated.

Multicollinearity and micronumerosity: what's the problem?

- We have relatively large standard errors with either:
  - Near multicollinearity, which is "high" but imperfect. In other words, one covariate is predicted very well by the others, but not perfectly.
  - Near micronumerosity, which means the number of observations barely exceeds the number of parameters to be estimated.
- OLS is still BLUE under multicollinearity or micronumerosity, though. These issues have nothing to do with statistical assumptions, they're just features of the data.

Diagnosing multicollinearity

- Telltale sign: ballooning standard errors, even though good model fit.
- The variance inflation factor (VIF).
- VIF for a given variable j: $\frac{1}{1-R_j^2}$, where R_j^2 is simply the R^2 value we obtain when regressing variable j on the other independent variables.
- Perfect VIF = 1, anything over 10 is cause for concern.

Diagnosing multicollinearity

VIF

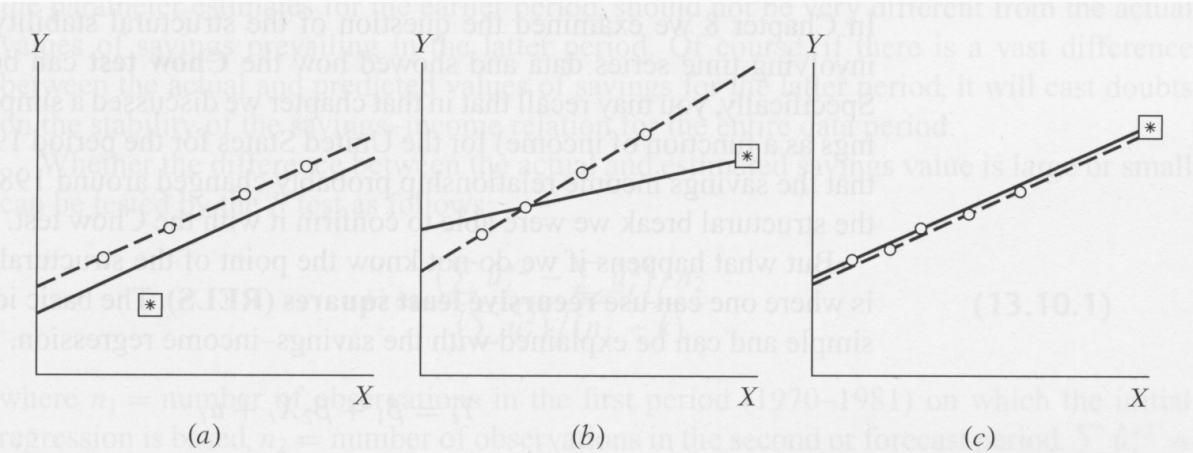
```
ols1 <- lm(Life.Exp ~ Illiteracy + Murder + HS.Grad +
Frost, data=states)
summary(ols1)
vif(ols1)
```

Dealing with multicollinearity and micronumerosity

0. Do nothing.
1. Dropping variables and specification bias.
2. Additional or new data.
3. Introducing polynomial terms (multicollinearity).
4. Measurement models: factor analysis, PCA, other data reduction techniques.

Outliers, leverage, and influence points

- An **outlier** is any observation that has a large residual. (a)
- A **leverage point** is any observation that is disproportionately distant from the bulk of the values of a regressor(s). (c)
- An **influential point** is an observation that is far from the bulk of the values of a regressor *AND* pulls the regression line towards itself. (b)



Source: Adapted from John Fox, op. cit., p. 268.

The `states` dataset

Studentized residuals

```
library(car)
ols1 <- lm(HS.Grad ~ Murder + Life.Exp + Frost, data=states)
summary(ols1)
qqPlot(ols1)
```

The `states` dataset

- The hat matrix H is defined as $X(X'X)^{-1}X'$
- Hence, H maps between the observed and predicted values of y ($X\hat{\beta}$).
- We adjust H when calculating robust standard errors to deal with heteroskedasticity.

Leverage: hat-values

```
influenceIndexPlot(ols1, vars="hat")
states["Hawaii",]
```

The `states` dataset

- Recall that influential observations are those that are both outliers *and* have high leverage.
- Measured with Cook's distance (D_i) which weighs both residuals and leverage (the hat-values).
- The size of the circles are proportional to Cook's D_i:

Influential observations

```
influencePlot(ols1)
#
ols2 <- update(ols1, subset=rownames(states)!="Nevada")
compareCoefs(ols1, ols2)
states["Nevada",]
mean(states$HS.Grad)
```

Addressing problematic data

Solutions

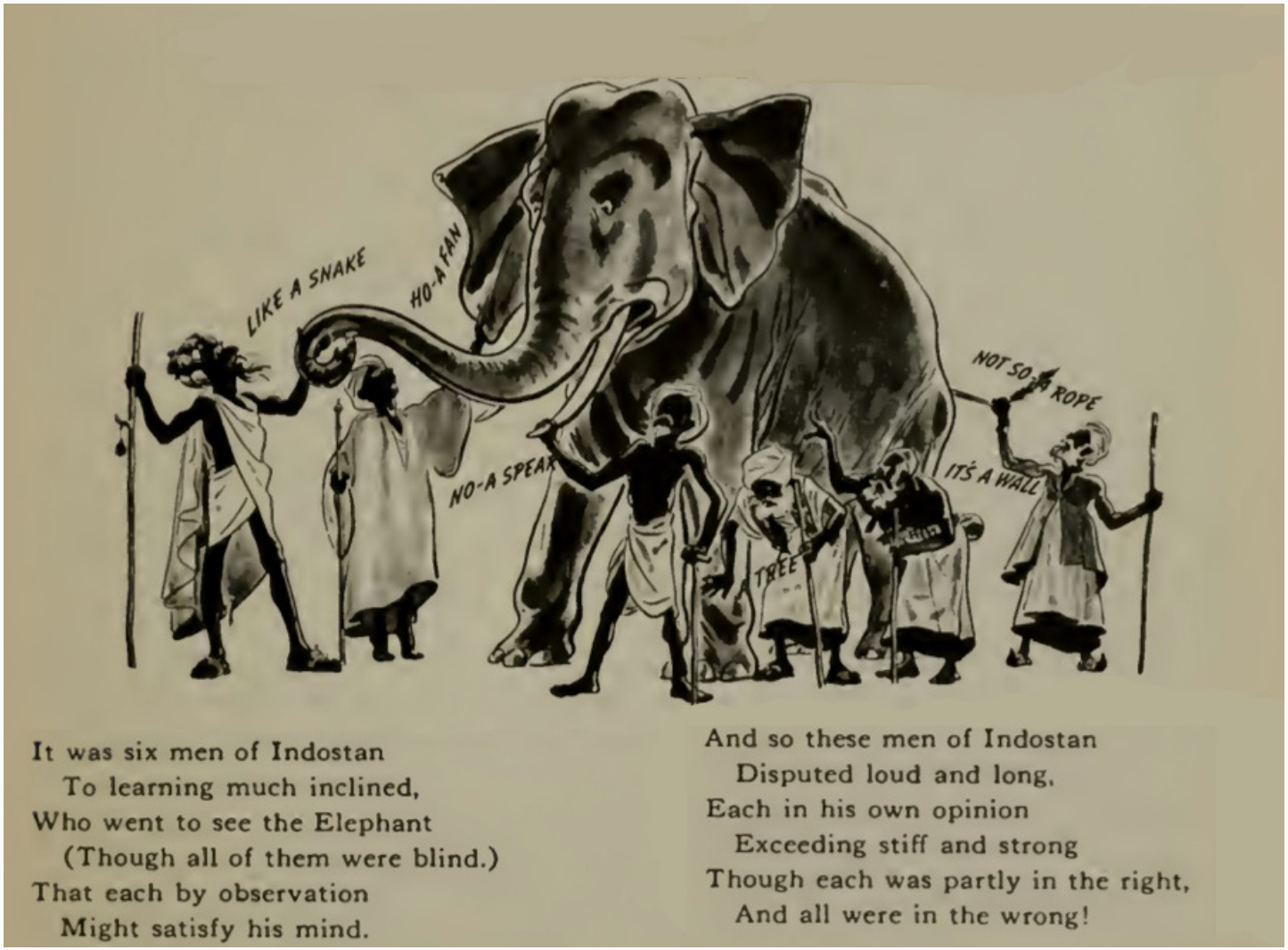
- One option is to remove problematic observations and re-estimate the model. Be very cautious if you take this approach, though.
- Draper & Smith (1998): Outliers may convey information that other data cannot. Outliers may warrant careful investigation. Only when traced to recording error should outliers be rejected automatically.
- Can you identify *why* an observation is an outlier, leverage, or influence point? Such information can usually guide your decision.

Uncertainty and Assessment

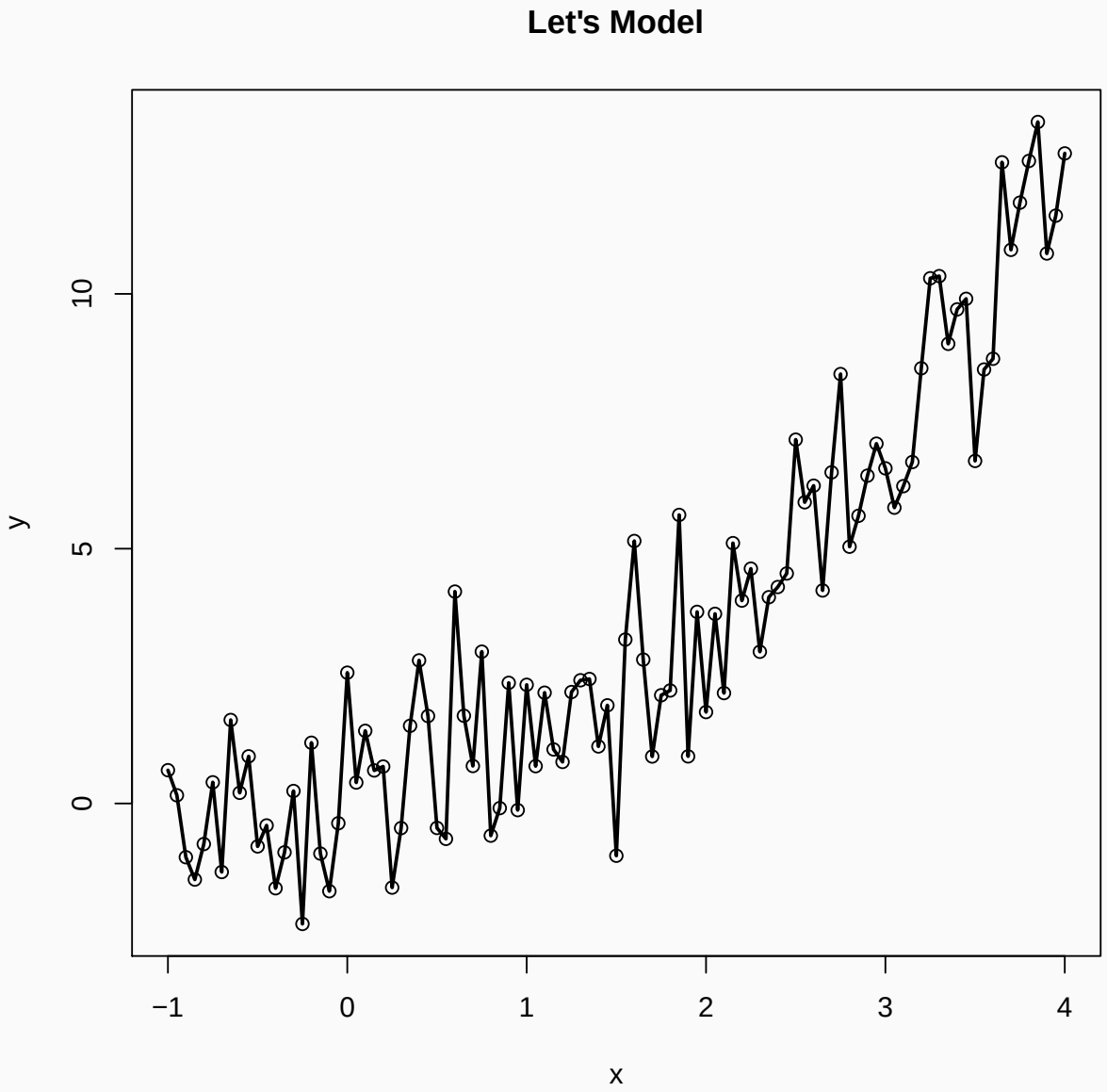
All models are wrong, but some are useful (George Box)

Three Musicians, 1921 by Pablo Picasso

The Rashomon effect



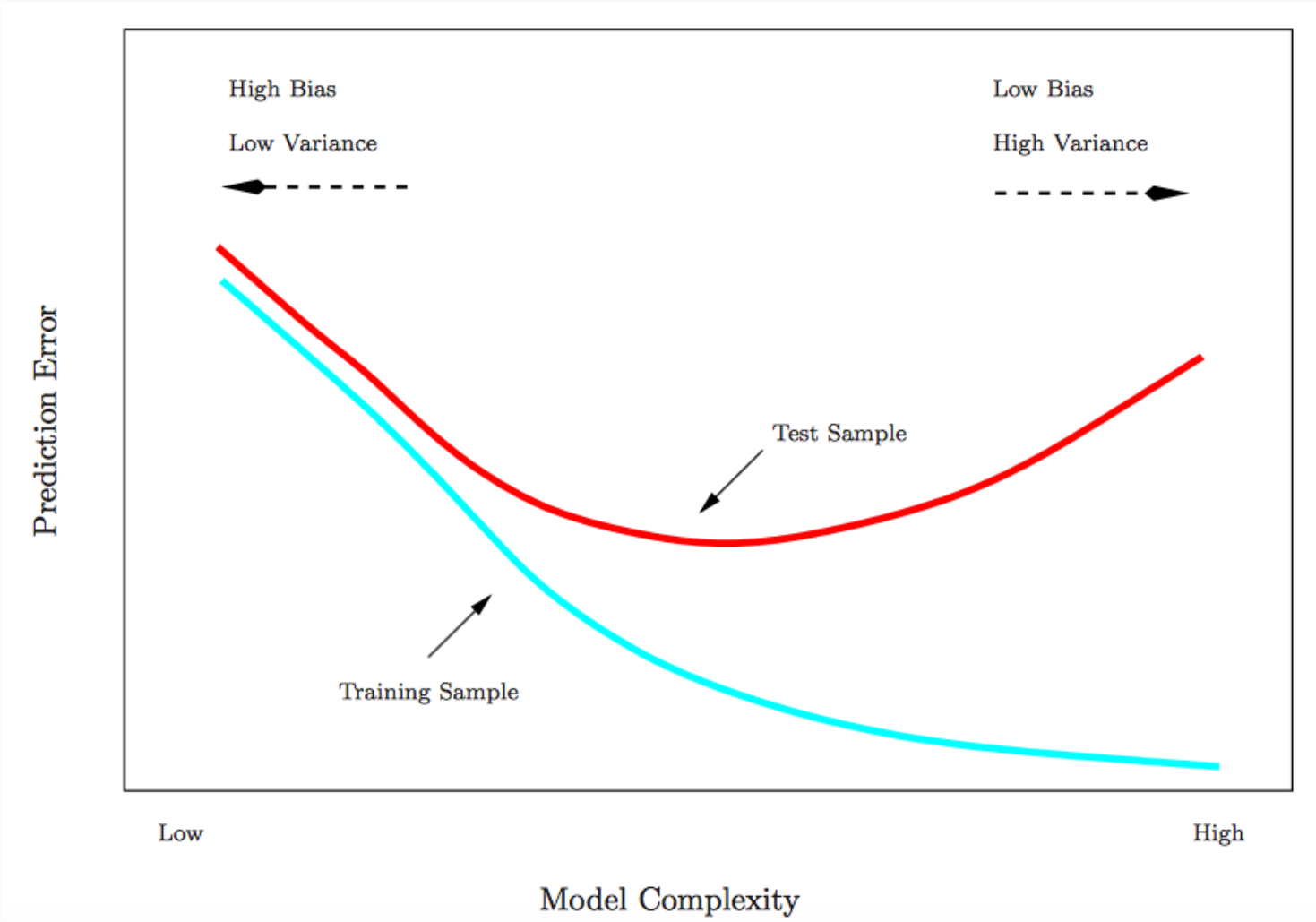
Now back to the bias-variance tradeoff. . .



Back to the bias-variance tradeoff...

- A fundamental challenge of data modeling is to strike the right balance between under and over-fitting the data, the **bias-variance tradeoff**.
- That is, when we're estimating f, we don't want to ignore true nonlinear complexities, but we also want parsimonious models of social/behavioral phenomena that *generalize* well.
- Our estimated function $\hat{f}$ should be an approximation to f (i.e., the true data-generating process) that minimizes bias and variance.

The bias-variance tradeoff



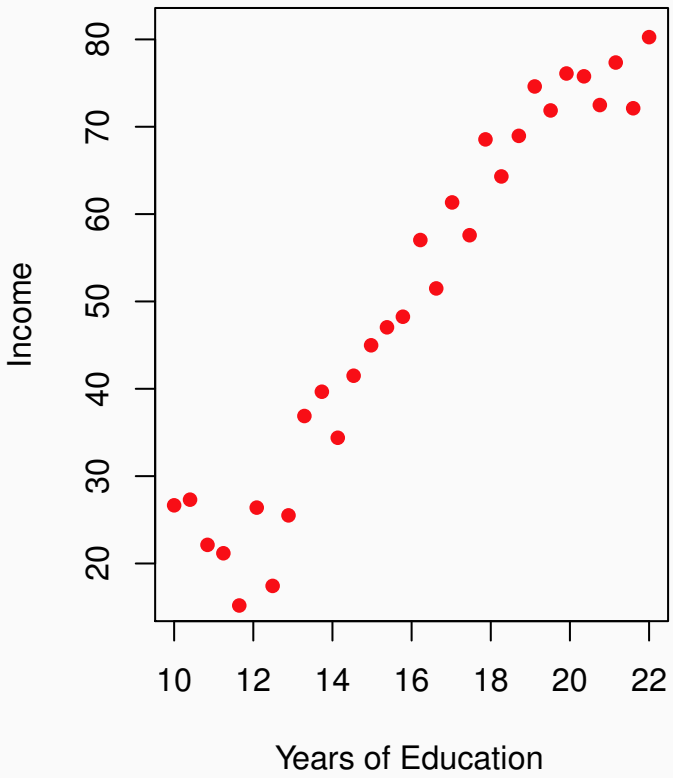
Model error

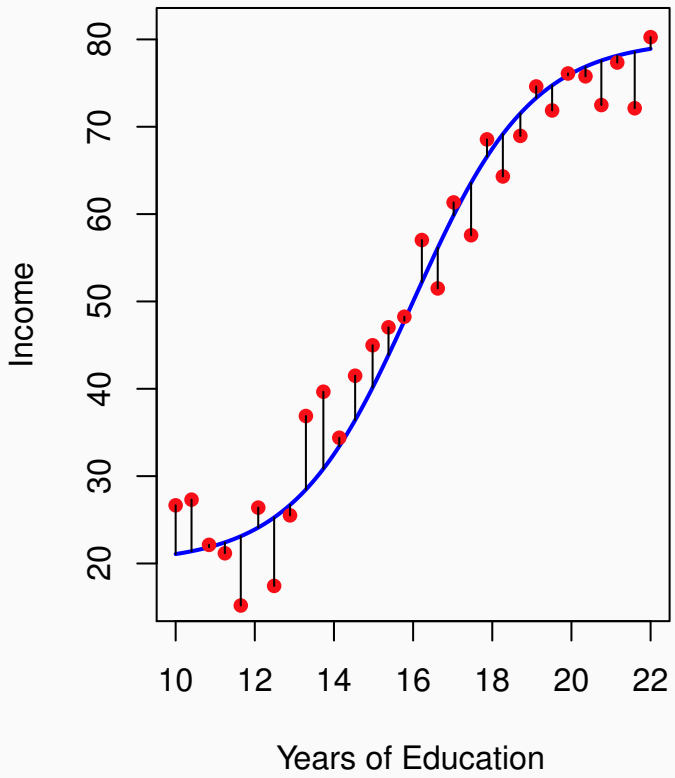
Model performance or fit is assessed with the use of some **loss function**, which aggregates model error (ε), or the discrepancy/residuals between actual values of Y and predicted values of Y (expressed as $\hat{Y}$). That is:

$$\hat{Y} = \hat{f}(X)$$
$$\varepsilon = Y - \hat{Y}$$

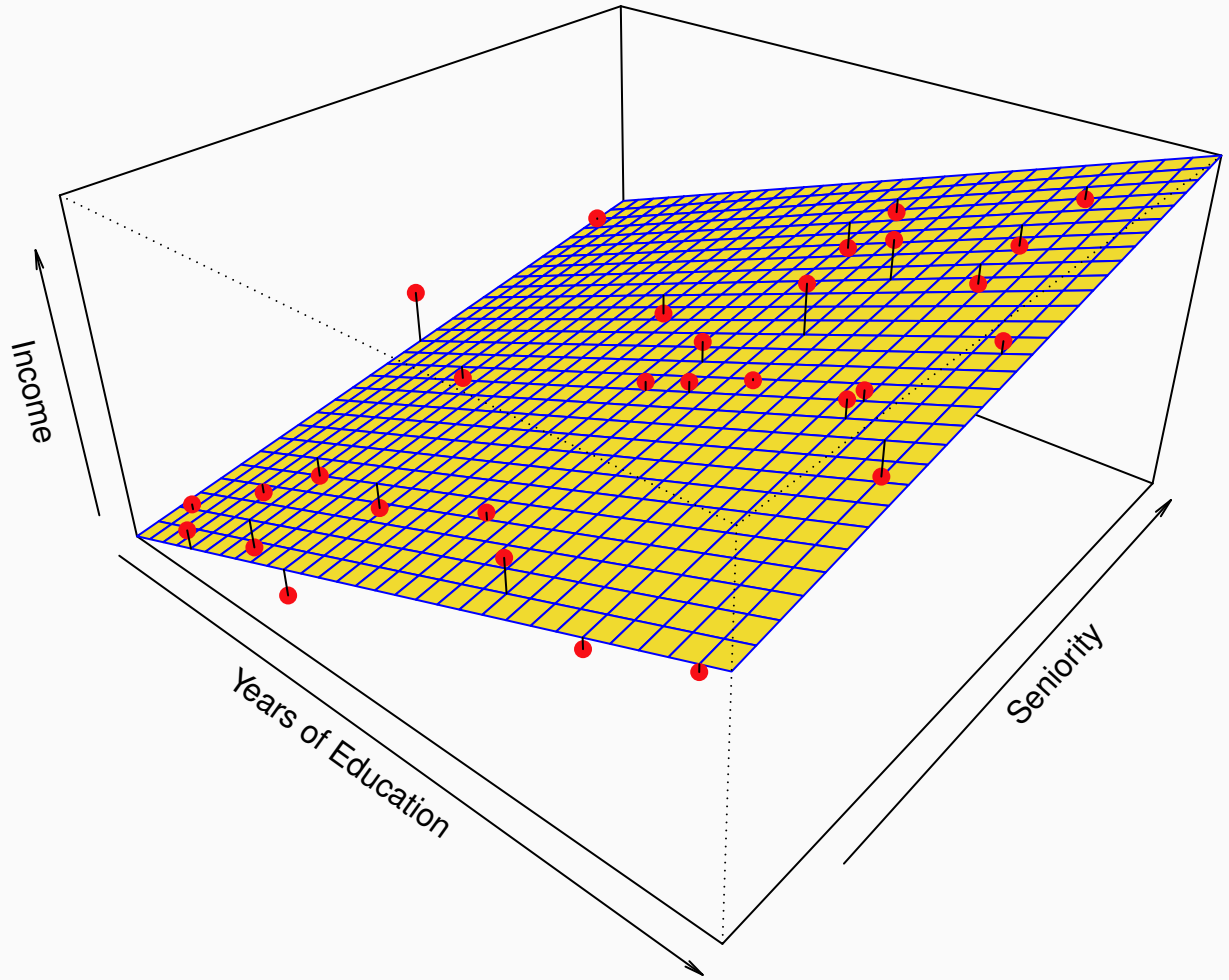
The OLS loss function, for instance, looks to minimize the sum of squared errors (SSE): $\sum \varepsilon^2$ or $\sum (Y - \hat{Y})^2$. You'll encounter other kinds of objective functions (e.g., the likelihood function), but the conceptual goal is the same: we want to bring the model into alignment with the data.

Residuals (with single predictor)





Residuals (with multiple predictors)



Model error

It is helpful to decompose ε (the difference between Y and $\hat{Y}$) into two components:

1. **Reducible error**: error due to disparities between $\hat{f}$ and f.
2. **Irreducible error**: error due to stochastic elements that is built into Y, separate from the data-generating process of f.
    - That is, $Y = f(X) + \varepsilon$: there will still be error remaining even if we approximate f exactly.

Model error

Hence, we need estimates of both kinds of model error to deal with the bias-variance tradeoff:

1. The error rate for the data used to estimate the model: how well does the model fit existing data?
2. The error rate for outside data: how well does the model fit new data?

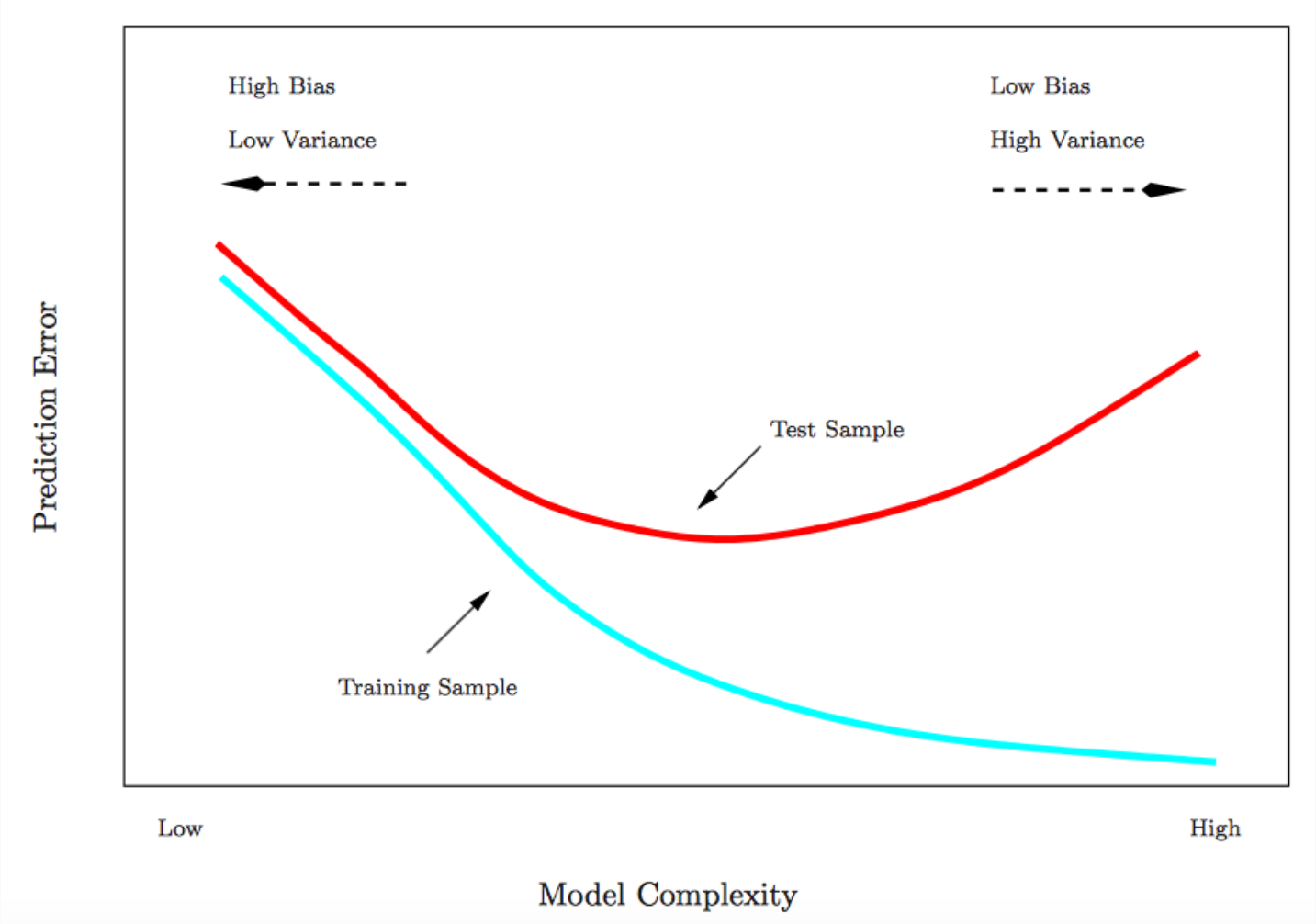
Training and testing error rates

Training error Error rate produced when the model is applied to **in-sample** data.

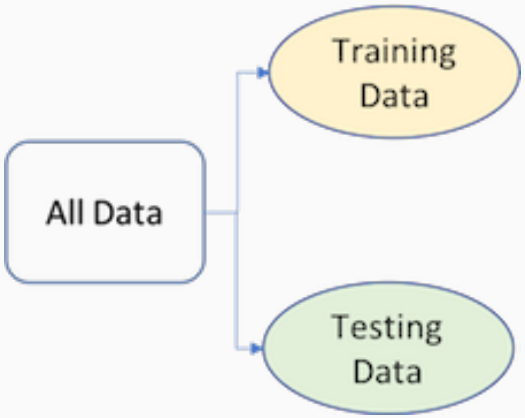
Testing error The average error that results from using a model to predict the response on a new, **out-of-sample** observation.

The two types of error are quite different: in particular, the training error rate can *drastically* understate the test error rate.

Remember me?


High Bias
Low Variance
Low Bias
High Variance
Test Sample
Training Sample
Prediction Error
Low
High
Model Complexity

Cross-validation

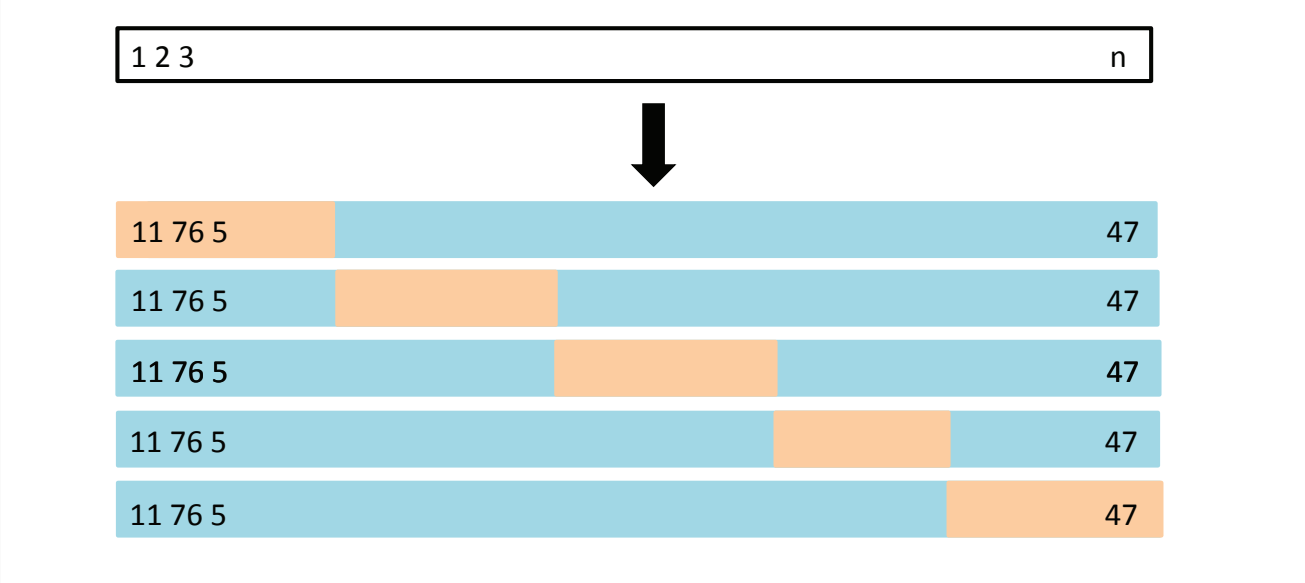


How do we estimate training and test error? A validation-set approach is one popular resampling technique:

1. Randomly divide the available set of samples into two parts: a training set and a test or hold-out set.
2. Estimate the model using the remaining of the data.
3. Apply the model to the observations in that subset, generating predictions ($\hat{Y}_{\text{test}} = \hat{f}(X_{\text{test}})$) and residuals ($Y_{\text{test}} - \hat{Y}_{\text{test}}$) to estimate testing error.

K-fold cross-validation

- Of course, we hate to lose data when estimating the model.
- K-fold cross-validation offers one solution: randomly divide the data into K equal-sized parts. We leave out part k, fit the model to the other $K-1$ parts (combined), and then obtain predictions for the left-out kth part.
- This is done in turn for each part $k = 1, 2, \ldots, K$ and then the results are combined.



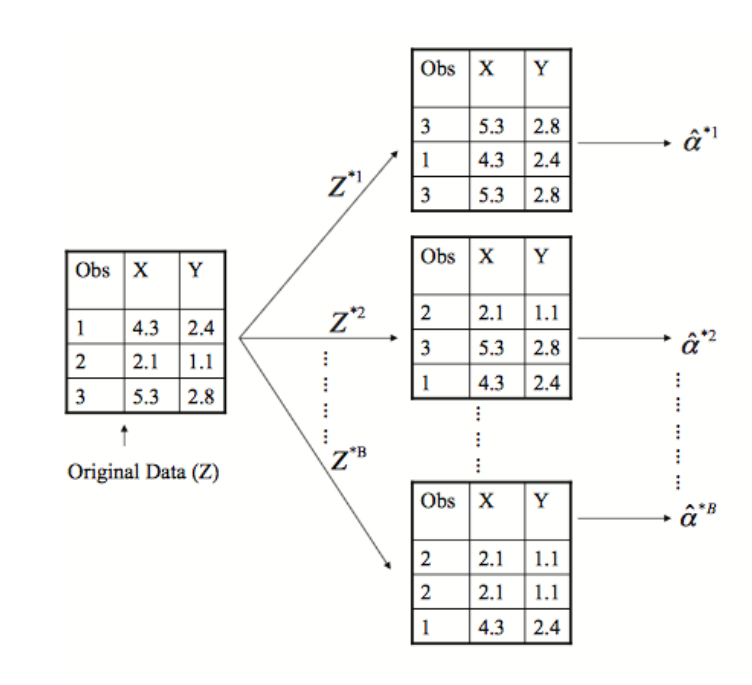
Measuring uncertainty

- What's the real source of our uncertainty about θ? Recall we just have $\hat{\theta}$ based on…?
- We want more of these. Can we use simulations to pull ourselves up by our bootstraps?

The bootstrap

- The **bootstrap** is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given statistic (e.g., the SE of a coefficient or a confidence interval for a difference-of-means test) using resampling.
- Quick aside: the generic terms *bootstrap* or *bootstrapping* usually refer to the nonparametric bootstrap, but there is are also a parametric bootstrap.

The bootstrap procedure



1. Randomly sample (with replacement) the original dataset to create a new sample.
2. Estimate the model on the new dataset, store estimates.
3. Repeat (let's say 100) times.
    - This gives us 100 point estimates of $\hat{\beta}$.
    - This is a nonparametric estimate of the sampling distribution of $\hat{\beta}$.
    - Can be used to give you a distribution of any statistic of interest in your model.

Why important?

- The bootstrap approach allows us to mimic the process of obtaining new data sets, so that we can estimate the variability and sensitivity of our estimates without the cost of generating additional samples.
- Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set with replacement.
- Each of these "bootstrap data sets" is created by sampling *with replacement*, and is the same size as our original dataset. As a result some observations may appear more than once in a given bootstrap data set and some not at all.

The bootstrap mantra

The population is to the sample

as

The sample is to the bootstrap sample

When to use the bootstrap

Anytime, but especially:

- In situations where ready-made standard errors are not available.
- Anytime that model assumptions involving standard errors are violated/likely to be violated:
  1. Normal errors
  2. You're worried about *N* and/or outliers
  3. Complex sampling designs with nonrandom selection of sample from the population (e.g., stratification/clustering/oversampling)

Tomz and van Houweling (2003) "How Does Voting Equipment Affect the Racial Gap in Voided Ballots?"

Can also use the bootstrap to conduct hypothesis testing.

- What is the probability that black Americans undervote at a higher rate than white Americans? Is there evidence of discrimination in voting equipment?
- Simulations can be used to estimate and/or incorporate uncertainty about θ.

Table 5: Intentional Undervoting by Race (NES, 1964-2000)

	Black		White	
	%	N	%	N
Voted for president	99.09	1,091	99.42	9,770
Didn't vote for pres	0.91	10	0.58	57

Tomz and van Houweling (2003) "How Does Voting Equipment Affect the Racial Gap in Voided Ballots?"

[18]For both racial groups the probability of intentional undervoting is close to zero. Under these conditions the normal approximation to the binomial distribution is poor, even with a large sample, making classical significance tests such as the chi-square inappropriate (Agresti 1992). We adopted a Bayesian approach. Specifically, drew random variates from two independent Beta distributions, $\pi_{black}|\text{data} \sim \text{beta}(11, 1092)$ and $\pi_{white}|\text{data} \sim \text{beta}(58, 9771)$, the posterior densities for black and white undervoting rates given the NES data and a uniform prior (Johnson and Albert 1999: 11). Drawing two numbers at a time, one from the black posterior and one from the white, we computed the Bayesian p-value as the relative frequency with which the draw of π_{white} exceeded the draw of π_{black}.

Drawing from sampling distributions

```
W <- rbeta(10000, 58, 9771)
B <- rbeta(10000, 11, 1092)
plot(density(B), lwd=2, ylim=c(0,600))
lines(density(W), lwd=2, lty=2)
table(B > W)
```