Week 4: Testing/Regression

Brandon Stewart\textsuperscript{1}

Princeton

October 3/5, 2016

\textsuperscript{1}These slides are heavily influenced by Matt Blackwell, Adam Glynn and Jens Hainmueller.
Where We’ve Been and Where We’re Going...

- **Last Week**
  - inference and estimator properties
  - point estimates, confidence intervals

- **This Week**
  - Monday:
    - hypothesis testing
    - what is regression?
  - Wednesday:
    - nonparametric regression
    - linear approximations

- **Next Week**
  - inference for simple regression
  - properties of OLS

- **Long Run**
  - probability → inference → regression

Questions?
Testing: Making Decisions
- Hypothesis testing
- Forming rejection regions
- P-values

Review: Steps of Hypothesis Testing

The Significance of Significance

Preview: What is Regression

Fun With Salmon

Nonparametric Regression
- Discrete X
- Continuous X
- Bias-Variance Tradeoff

Linear Regression
- Combining Linear Regression with Nonparametric Regression
- Least Squares

Interpreting Regression

Fun With Linearity
A Running Example for Testing

Statistics play an important role in determining which drugs are approved for sale by the FDA.

There are typically three phases of clinical trials before a drug is approved:

- Phase I: Toxicity (Will it kill you?)
- Phase II: Efficacy (Is there any evidence that it helps?)
- Phase III: Effectiveness (Is it better than existing treatments?)

Phase I trials are conducted on a small number of healthy volunteers, Phase II trial are either randomized experiments or within-patient comparisons, and Phase III trials are almost always randomized experiments with control groups.
Consider a Phase II efficacy trial reported in Sowers et al. (2006), for a drug combination designed to treat high blood pressure in patients with metabolic syndrome.

- The trial included 345 patients with initial systolic blood pressure between 140-159.
- Each subject was assigned to take the drug combination for 16 weeks.
- Systolic blood pressure was measured on each subject before and after the treatment period.
<table>
<thead>
<tr>
<th>Subject</th>
<th>SBP\textsubscript{before}</th>
<th>SBP\textsubscript{after}</th>
<th>Decrease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>147</td>
<td>135</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>153</td>
<td>122</td>
<td>31</td>
</tr>
<tr>
<td>3</td>
<td>142</td>
<td>119</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>141</td>
<td>134</td>
<td>7</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>345</td>
<td>155</td>
<td>115</td>
<td>40</td>
</tr>
</tbody>
</table>
Example

- The drug was administered to 345 patients.
- On average, blood pressure was 21 points lower after treatment.
- The standard deviation of changes in blood pressure was 14.3.

Question: Should the FDA allow the drug to proceed to the next stage of testing?
The FDA’s Decision

We can think of the FDA’s problem in terms of two dimensions:

- The true state of the world
- The decision made by the FDA

<table>
<thead>
<tr>
<th></th>
<th>Drug works</th>
<th>Drug doesn’t work</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDA approves</td>
<td>Good!</td>
<td>Bad!</td>
</tr>
<tr>
<td>FDA doesn’t approve</td>
<td>Bad!</td>
<td>Good!</td>
</tr>
</tbody>
</table>

Stewart (Princeton)
Testing: Making Decisions
- Hypothesis testing
- Forming rejection regions
- P-values

Review: Steps of Hypothesis Testing

The Significance of Significance

Preview: What is Regression

Fun With Salmon

Nonparametric Regression
- Discrete X
- Continuous X
- Bias-Variance Tradeoff

Linear Regression
- Combining Linear Regression with Nonparametric Regression
- Least Squares

Interpreting Regression

Fun With Linearity
Elements of a Hypothesis Test

Hypothesis testing gives us a systematic framework for making decisions based on observed data.

Terms to know:

- Null Hypothesis
- Alternative Hypothesis
- Test Statistic
- Rejection Region
Null and Alternative Hypotheses

- **Null Hypothesis**: The conservatively assumed state of the world (often “no effect”)
  
  Example: The drug does not reduce blood pressure on average \((\mu_{decrease} \leq 0)\)

- **Alternative Hypothesis**: Claim to be tested (research hypothesis)
  
  Example: The drug does reduce blood pressure on average \((\mu_{decrease} > 0)\)
More Examples

Null Hypothesis Examples ($H_0$):

- The drug does not change blood pressure on average ($\mu_{decrease} = 0$)

Alternative Hypothesis Examples ($H_a$):

- The drug does change blood pressure on average ($\mu_{decrease} \neq 0$)
The FDA’s Decision

Back to the two dimensions of the FDA’s problem:
- The true state of the world
- The decision made by the FDA

<table>
<thead>
<tr>
<th>Drug works ($H_0$ False)</th>
<th>Drug doesn’t work ($H_0$ True)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDA approves (reject $H_0$)</td>
<td>Correct</td>
</tr>
<tr>
<td>FDA doesn’t approve (don’t reject $H_0$)</td>
<td>Type II error</td>
</tr>
</tbody>
</table>
Test Statistic: A function of the sample and the null hypothesis value of the parameter. For example:

\[ \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \]

Null Distribution: the sampling distribution of the statistic/test statistic assuming that the null is true.
Null Distributions

The CLT tells us that in large samples,

$$\overline{X} \sim_{approx} N(\mu, \sigma^2/n).$$

We know from our previous discussion that in large samples,

$$S/\sqrt{n} \approx \sigma/\sqrt{n}$$

If we assume that the null hypothesis is true such that $\mu = \mu_0$, then

$$\overline{X} \sim_{approx} N(\mu_0, S^2/n)$$

$$\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim_{approx} N(0, 1)$$
1 Testing: Making Decisions
- Hypothesis testing
- Forming rejection regions
- P-values

2 Review: Steps of Hypothesis Testing

3 The Significance of Significance

4 Preview: What is Regression

5 Fun With Salmon

6 Nonparametric Regression
- Discrete X
- Continuous X
- Bias-Variance Tradeoff

7 Linear Regression
- Combining Linear Regression with Nonparametric Regression
- Least Squares

8 Interpreting Regression

9 Fun With Linearity
$\alpha$ is the probability of Type I error.

We usually pick an $\alpha$ that we are comfortable with in advance, and using the null distribution for the test statistic and the alternative hypothesis, we define a rejection region.

Example: Suppose $\alpha = 5\%$, the test statistic is $\frac{\bar{X} - \mu_0}{S/\sqrt{n}}$, the null hypothesis is $H_0 : \mu = \mu_0$, and the alternative hypothesis is $H_a : \mu \neq \mu_0$. 
Two-sided rejection region

Rejection region with $\alpha = .05, H_0 : \mu = 0, H_A : \mu \neq 0$:
One-sided Rejection Region

Rejection region with $\alpha = .05$, $H_0 : \mu \leq 0$, $H_A : \mu > 0$:

![Graph showing one-sided rejection region with critical value shaded.](image-url)
Example

So, should the FDA approve further trials?

Recall the null and alternative hypotheses:

\[ H_0 : \mu_{\text{decrease}} \leq 0 \]

\[ H_a : \mu_{\text{decrease}} > 0 \]
We can calculate the test statistic:

- $\bar{x} = 21.0$
- $s = 14.3$
- $n = 345$

Therefore,

$$T = \frac{21.0 - 0}{\frac{14.3}{\sqrt{345}}} = 27.3$$

What is the decision?
Rejection Region with $\alpha = .05$
Rejection Region with $\alpha = .05$
Testing: Making Decisions
- Hypothesis testing
- Forming rejection regions
- P-values

Review: Steps of Hypothesis Testing

The Significance of Significance

Preview: What is Regression

Fun With Salmon

Nonparametric Regression
- Discrete X
- Continuous X
- Bias-Variance Tradeoff

Linear Regression
- Combining Linear Regression with Nonparametric Regression
- Least Squares

Interpreting Regression

Fun With Linearity
P-value

The appropriate level ($\alpha$) for a hypothesis test depends on the relative costs of Type I and Type II errors.

What if there is disagreement about these costs?

We might like a quantity that summarizes the strength of evidence against the null hypothesis without making a yes or no decision.

**P-value**: Assuming that the null hypothesis is true, the probability of getting something at least as extreme as our observed test statistic, where extreme is defined in terms of the alternative hypothesis.
P-value

The p-value depends on both the realized value of the test statistic and the alternative hypothesis.

\[
H_a : \mu > 0 \quad \quad \quad H_a : \mu \neq 0 \quad \quad \quad H_a : \mu < 0
\]

\[
p = 0.036 \quad \quad \quad p = 0.072 \quad \quad \quad p = 0.964
\]
Rejection Regions and P-values

What is the relationship between p-values and the rejection region of a test? Assume that $\alpha = .05$:

$$H_a : \mu > 0$$

$$H_a : \mu \neq 0$$

$$H_a : \mu < 0$$

If $p < \alpha$, then the test statistic falls in the rejection region for the $\alpha$-level test.
Example 1

Recall the drug testing example, where $H_0 : \mu_0 \leq 0$ and $H_a : \mu_0 > 0$:

- $\bar{x} = 21.0$
- $s = 14.3$
- $n = 345$

Therefore,

$$T = \frac{21.0 - 0}{\frac{14.3}{\sqrt{345}}} = 27.3$$

What is the probability of observing a test statistic greater than 27.3 if the null is true?
Example 1

Test Statistic

Fail to Reject

Reject

$p < .0001$

$T = 27.3$
Rejection Regions and $1 - \alpha$ CIs

Up to this point, we have defined rejection regions in terms of the test statistic.

In some cases, we can define an equivalent rejection region in terms of the parameter of interest.

For a two-sided, large-sample test, we reject if:

\[
\frac{\bar{X} - \mu_0}{s/\sqrt{n}} > z_{\alpha/2} \quad \text{or} \quad \frac{\bar{X} - \mu_0}{s/\sqrt{n}} < -z_{\alpha/2}
\]

\[
\bar{X} - \mu_0 > z_{\alpha/2} \times \frac{s}{\sqrt{n}} \quad \text{or} \quad \bar{X} - \mu_0 < -z_{\alpha/2} \times \frac{s}{\sqrt{n}}
\]

\[
\bar{X} > \mu_0 + z_{\alpha/2} \times \frac{s}{\sqrt{n}} \quad \text{or} \quad \bar{X} < \mu_0 - z_{\alpha/2} \times \frac{s}{\sqrt{n}}
\]
The rescaled rejection region is related to $1 - \alpha$ CI:

- If the observed $\bar{X}$ is in the $\alpha$ rejection region, the $1 - \alpha$ CI does not contain $\mu_0$.
- If the observed $\bar{X}$ is not in the $\alpha$ rejection region, the $1 - \alpha$ CI contains $\mu_0$.

Therefore, we can use the $1 - \alpha$ CI to test the null hypothesis at the $\alpha$ level.
Another interpretation of CIs

The form of the “fail to reject” region of an $\alpha$-level hypothesis test is:

$$
\left( \mu_0 - z_{\alpha/2} \times \frac{s}{\sqrt{n}}, \mu_0 + z_{\alpha/2} \times \frac{s}{\sqrt{n}} \right)
$$

The form of a region of a $1 - \alpha$ CI is:

$$
\left( \bar{X} - z_{\alpha/2} \times \frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \times \frac{s}{\sqrt{n}} \right)
$$

So the $1 - \alpha$ CI is the set of null hypotheses $\mu_0$ that would not be rejected at the $\alpha$ level.
Testing: Making Decisions
- Hypothesis testing
- Forming rejection regions
- P-values

Review: Steps of Hypothesis Testing

The Significance of Significance

Preview: What is Regression

Fun With Salmon

Nonparametric Regression
- Discrete X
- Continuous X
- Bias-Variance Tradeoff

Linear Regression
- Combining Linear Regression with Nonparametric Regression
- Least Squares

Interpreting Regression

Fun With Linearity
Hypothesis Testing: Setup

Goal: test a hypothesis about the value of a parameter.

Statistical decision theory underlies such hypothesis testing.

Trial Example:

Suppose we must decide whether to convict or acquit a defendant based on evidence presented at a trial. There are four possible outcomes:

<table>
<thead>
<tr>
<th>Decision</th>
<th>Defendant</th>
<th>Guilty</th>
<th>Innocent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convict</td>
<td>Correct</td>
<td>Type I Error</td>
<td></td>
</tr>
<tr>
<td>Acquit</td>
<td>Type II Error</td>
<td>Correct</td>
<td></td>
</tr>
</tbody>
</table>

We could make two types of errors:

- Convict an innocent defendant (type-I error)
- Acquit a guilty defendant (type-II error)

Our goal is to limit the probability of making these types of errors.

However, creating a decision rule which minimizes both types of errors at the same time is impossible. We therefore need to balance them.
### Hypothesis Testing: Error Types

<table>
<thead>
<tr>
<th>Decision</th>
<th>Defendant</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Guilty</td>
<td>Acquit</td>
<td>Correct</td>
</tr>
<tr>
<td>Convict</td>
<td></td>
<td>Type-II error</td>
<td>Correct</td>
</tr>
<tr>
<td>Acquit</td>
<td>Correct</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type-I error</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Now, suppose that we have a statistical model for the probability of convicting and acquitting, conditional on whether the defendant is actually guilty or innocent.

Then, our decision-making rule can be characterized by two probabilities:

- $\alpha = \Pr(\text{type-I error}) = \Pr(\text{convict} \mid \text{innocent})$
- $\beta = \Pr(\text{type-II error}) = \Pr(\text{acquit} \mid \text{guilty})$

The probability of making a correct decision is therefore $1 - \alpha$ (if innocent) and $1 - \beta$ (if guilty).

Hypothesis testing follows an analogous logic, where we want to decide whether to reject (= convict) or fail to reject (= acquit) a null hypothesis (= defendant) using sample data.
Hypothesis Testing: Steps

<table>
<thead>
<tr>
<th>Decision</th>
<th>Null Hypothesis ($H_0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>False</td>
</tr>
<tr>
<td>Reject</td>
<td>$1 - \beta$</td>
</tr>
<tr>
<td>Fail to Reject</td>
<td>$\beta$</td>
</tr>
</tbody>
</table>

1. Specify a **null hypothesis** $H_0$ (e.g. the defendant = innocent)

2. Pick a value of $\alpha = \Pr(\text{reject } H_0 \mid H_0)$ (e.g. 0.05). This is the maximum probability of making a type-I error we decide to tolerate, and called the **significance level** of the test.

3. Choose a **test statistic** $T$, which is a function of sample data and related to $H_0$ (e.g. the count of testimonies against the defendant)

4. Assuming $H_0$ is true, derive the **null distribution** of $T$ (e.g. standard normal)
Hypothesis Testing: Steps

<table>
<thead>
<tr>
<th>Null Hypothesis ($H_0$)</th>
<th>False</th>
<th>True</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision</td>
<td>Reject</td>
<td>$1 - \beta$</td>
</tr>
<tr>
<td></td>
<td>Fail to Reject</td>
<td>$\beta$</td>
</tr>
</tbody>
</table>

Using the **critical values** from a statistical table, evaluate how unusual the observed value of $T$ is under the null hypothesis:

- If the probability of drawing a $T$ at least as extreme as the observed $T$ is less than $\alpha$, we reject $H_0$.
  (e.g. there are too many testimonies against the defendant for her to be innocent, so reject the hypothesis that she was innocent.)

- Otherwise, we fail to reject $H_0$.
  (e.g. there is not enough evidence against the defendant, so give her the benefit of the doubt.)
Testing: Making Decisions
- Hypothesis testing
- Forming rejection regions
- P-values

Review: Steps of Hypothesis Testing

The Significance of Significance

Preview: What is Regression

Fun With Salmon

Nonparametric Regression
- Discrete X
- Continuous X
- Bias-Variance Tradeoff

Linear Regression
- Combining Linear Regression with Nonparametric Regression
- Least Squares

Interpreting Regression

Fun With Linearity
Practical versus Statistical Significance

\[ \frac{X - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \]

- What are the possible reasons for rejecting the null?
  1. \( X - \mu_0 \) is large (big difference between sample mean and mean assumed by \( H_0 \))
  2. \( n \) is large (you have a lot of data so you have a lot of precision)
  3. \( S \) is small (the outcome has low variability)

- We need to be careful to distinguish:
  - practical significance (e.g. a big effect)
  - statistical significance (i.e. we reject the null)

- In large samples even tiny effects will be significant, but the results may not be very important substantively. Always discuss both!

Stewart (Princeton)  
Week 4: Testing/Regression  
October 3/5, 2016
Star Chasing (aka there is an XKCD for everything)
Multiple Testing

- If we test all of the coefficients separately with a t-test, then we should expect that 5% of them will be significant just due to random chance.
- Illustration: randomly draw 21 variables, and run a regression of the first variable on the rest.
- By design, no effect of any variable on any other, but when we run the regression:
Multiple Test Example

```
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.02804 0.11382 -0.246 0.80605
## X2 -0.15039 0.11218 -1.341 0.18389
## X3 0.07916 0.09503 0.833 0.40736
## X4 -0.07174 0.10458 -0.686 0.49472
## X5 0.17208 0.11400 1.509 0.13518
## X6 0.08085 0.10834 0.746 0.45772
## X7 0.10292 0.11416 0.902 0.37006
## X8 -0.32105 0.12067 -2.661 0.00945 **
## X9 -0.05312 0.10798 -0.492 0.62412
## X10 0.18010 0.12644 1.424 0.15827
## X11 0.16639 0.11095 1.500 0.13768
## X12 0.00801 0.10377 0.077 0.93866
## X13 0.00021 0.10378 0.002 0.99838
## X14 -0.06597 0.11215 -0.588 0.55829
## X15 -0.12965 0.11416 -1.162 0.24872
## X16 -0.05445 0.12514 -0.435 0.66469
## X17 0.00433 0.11201 0.039 0.96923
## X18 -0.08079 0.10985 -0.735 0.46421
## X19 -0.08581 0.11855 -0.724 0.47134
## X20 -0.18601 0.10456 -1.779 0.07910 .
## X21 0.00211 0.10812 0.020 0.98447
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 0.9992 on 79 degrees of freedom
## Multiple R-squared:  0.2009, Adjusted R-squared:  -0.00142
## F-statistic: 0.993 on 20 and 79 DF,  p-value: 0.4797
```
Multiple Testing Gives False Positives

- Notice that out of 20 variables, one of the variables is significant at the 0.05 level (in fact, at the 0.01 level).
- But this is exactly what we expect: $1/20 = 0.05$ of the tests are false positives at the 0.05 level.
- Also note that $2/20 = 0.1$ are significant at the 0.1 level. Totally expected!
Problem of Multiple Testing

- The multiple testing (or “multiple comparison”) problem occurs when one considers a set of statistical tests simultaneously.
- Consider $k = 1, \ldots, m$ independent hypothesis tests (e.g. control versus various treatment groups). Even if each test is carried out at a low significance level (e.g., $\alpha = 0.05$) the overall type I error rate grows very fast:
  \[
  \alpha_{overall} = 1 - (1 - \alpha_k)^m.
  \]
- That’s right - it grows exponentially. E.g., given test 7 tests at $\alpha = .1$ level the overall type I error is .52.
- Even if all null hypotheses are true we will reject at least one of them with probability .52.
- Same for confidence intervals: probability that all 7 CI cover the true values simultaneously over repeated samples is .52. So for each coefficient you have a .90 confidence interval, but overall a .52 percent confidence interval.
Problem of Multiple Testing

- Several statistical techniques have been developed to “adjust” for this inflation of overall type I errors for multiple testing.
- To compensate for the number of tests, these corrections generally require a stronger level of evidence to be observed in order for an individual comparison to be deemed “significant”
- The most prominent adjustments include:
  - Bonferroni: for each individual test use significance level of
    \[ \alpha_{k,BFer} = \frac{\alpha_k}{m} \]
  - Sidak: for each individual test use significance level of
    \[ \alpha_{k,Sid} = 1 - (1 - \alpha_k)\frac{1}{m} \]
  - Scheffe (for confidence intervals)
- It remains a heated debate.
Summary of Testing

- Hypothesis testing provides a principled framework for making decisions between alternatives.
- The level of a test determines how often the researcher is willing to reject a correct null hypothesis.
- Reporting p-values allows the researcher to separate the analysis from the decision.
- There is a close relationship between the results of an $\alpha$ level hypothesis test and the coverage of a $(1 - \alpha)\%$ confidence interval.
What we’ve been up to: estimating parameters of population distributions. Generally we’ve been learning about a single variable.

From here on out, we’ll be interested in the relationships between variables. How does one variable change as we change the values of another variable? This question will be the bread and butter of the class moving forward.
What is a relationship and why do we care?

- Most of what we want to do in the social science is learn about how two variables are related
- Examples:
  - Does turnout vary by types of mailers received?
  - Is the quality of political institutions related to average incomes?
  - Does parental incarceration affect intergenerational mobility for child?
Notation and conventions

- \( Y \) - the dependent variable or outcome or regressand or left-hand-side variable or response
  - Voter turnout
  - Log GDP per capita
  - Income relative to parent

- \( X \) - the independent variable or explanatory variable or regressor or right-hand-side variable or treatment or predictor
  - Social pressure mailer versus Civic Duty Mailer
  - Average Expropriation Risk
  - Incarcerated parent

- Generally our goal is to understand how \( Y \) varies as a function of \( X \):

\[
Y = f(X) + \text{error}
\]
Three uses of regression

1. **Description** - parsimonious summary of the data
2. **Prediction/Estimation/Inference** - learn about parameters of the joint distribution of the data
3. **Causal Inference** - evaluate counterfactuals
Describing relationships

- Remember that we had ways to summarize the relationship between variables in the population.
- Joint densities, covariance, and correlation were all ways to summarize the relationship between two variables.
- But these were population quantities and we only have samples, so we may want to estimate these quantities using their sample analogs (plug-in principle or analogy principle).
Scatterplots

- Sample version of joint probability density.
- Shows graphically how two variables are related

Data from Acemoglu, Johnson and Robinson
Non-linear relationship

- Example of a non-linear relationship, where we use the unlogged version of GDP and settler mortality:
The sample version of population covariance, \( \sigma_{XY} = E[(X - E[X])(Y - E[Y])] \).

**Definition (Sample Covariance)**

The sample covariance between \( Y_i \) and \( X_i \) is

\[
S_{XY} = \frac{1}{n - 1} \sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)
\]
Sample Correlation

The sample version of population correlation, \( \rho = \sigma_{XY} / \sigma_X \sigma_Y \).

**Definition (Sample Correlation)**

The **sample correlation** between \( Y_i \) and \( X_i \) is

\[
\hat{\rho} = r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}}
\]
Regression is About Conditioning on $X$

- Regression quantifies how an outcome variable $Y$ varies as a function of one or more predictor variables $X$.
- Many methods, but the common idea: conditioning on $X$.
- Goal is to characterize $f(Y|X)$, the conditional probability distribution of $Y$ for different levels of $X$.
- Instead of modeling the whole conditional density of $Y$ given $X$, in regression we usually only model the conditional mean of $Y$ given $X$: $E[Y|X = x]$.
- Our key goal is to approximate the conditional expectation function $E[Y|X]$, which summarizes how the average of $Y$ varies across all possible levels of $X$ (also called the population regression function).
- Once we have estimated $E[Y|X]$, we can use it for prediction and/or causal inference, depending on what assumptions we are willing to make.
Review: Conditional expectation

It will be helpful to review a core concept:

**Definition (Conditional Expectation Function)**

The *conditional expectation function* (CEF) or the *regression function* of $Y$ given $X$, denoted

$$r(x) = E[Y|X = x]$$

is the function that gives the mean of $Y$ at various values of $x$.

- Note that this is a function of the population distributions. We will want to produce estimates $\hat{r}(x)$. 
CEF for binary covariates

- We’ve been writing $\mu_1$ and $\mu_0$ for the means in different groups.
- For example, on the homework, you are looking at the expected value of the loan amount conditional on gender. There we had $\mu_m$ and $\mu_w$.
- Note that these are just conditional expectations. Define $Y$ to be the loan amount, $X = 1$ to indicate a man, and $X = 0$ to indicate a woman and then we have:

$$\mu_m = r(1) = E[Y|X = 1]$$
$$\mu_w = r(0) = E[Y|X = 0]$$

- Notice here that since $X$ can only take on two values, 0 and 1, then these two conditional means completely summarize the CEF.
Estimating the CEF for binary covariates

- How do we estimate \( \hat{r}(x) \)?
- We’ve already done this: it’s just the usual sample mean among the men and then the usual sample mean among the women:

\[
\hat{r}(1) = \frac{1}{n_1} \sum_{i: X_i = 1} Y_i \\
\hat{r}(0) = \frac{1}{n_0} \sum_{i: X_i = 0} Y_i
\]

Here we have \( n_1 = \sum_{i=1}^n X_i \) is the number of men in the sample and \( n_0 = n - n_1 \) is the number of women.

- The sum here \( \sum_{i: X_i = 1} \) is just summing only over the observations \( i \) such that have \( X_i = 1 \), meaning that \( i \) is a man.

- This is very straightforward: estimate the mean of \( Y \) conditional on \( X \) by just estimating the means within each group of \( X \).
Binary covariate example CEF plot
The conditional expectation function $E[Y|X]$ is the \textbf{estimand} (or parameter) we are interested in.

$\hat{E}[Y|X]$ is the \textbf{estimator} of this parameter of interest, which is a function of $X$.

For a given sample dataset, we obtain an \textbf{estimate} of $E[Y|X]$.

We want to extend the regression idea to the case of multiple $X$ variables, but we will start this week with the simple bivariate case where we have a single $X$. 
Testing: Making Decisions
- Hypothesis testing
- Forming rejection regions
- P-values

Review: Steps of Hypothesis Testing

The Significance of Significance

Preview: What is Regression

Fun With Salmon

Nonparametric Regression
- Discrete X
- Continuous X
- Bias-Variance Tradeoff

Linear Regression
- Combining Linear Regression with Nonparametric Regression
- Least Squares

Interpreting Regression

Fun With Linearity
Fun With Salmon

Methods

(a.k.a. the greatest methods section of all time)

- **Subject**
  “One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.”

- **Task**
  “The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.”

- **Design**
  “Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.”
“Several active voxels were discovered in a cluster located within the salmon’s brain cavity. The size of this cluster was $81 \text{ mm}^3$ with a cluster-level significance of $p = .001$.\"
Where We’ve Been and Where We’re Going...

- **Last Week**
  - inference and estimator properties
  - point estimates, confidence intervals

- **This Week**
  - Monday:
    - hypothesis testing
    - what is regression?
  - Wednesday:
    - nonparametric regression
    - linear approximations

- **Next Week**
  - inference for simple regression
  - properties of OLS

- **Long Run**
  - probability $\rightarrow$ inference $\rightarrow$ regression

Questions?
Testing: Making Decisions
- Hypothesis testing
- Forming rejection regions
- P-values

Review: Steps of Hypothesis Testing

The Significance of Significance

Preview: What is Regression

Fun With Salmon

Nonparametric Regression
- Discrete X
- Continuous X
- Bias-Variance Tradeoff

Linear Regression
- Combining Linear Regression with Nonparametric Regression
- Least Squares

Interpreting Regression

Fun With Linearity
Nonparametric Regression with Discrete $X$

- Let’s take a look at some data on education and income from the American National Election Study.

- We use two variables:
  - $Y$: income
  - $X$: educational attainment

- Goal is to characterize the conditional expectation $E[Y|X]$, i.e. how average income varies with education level.
**Nonparametric Regression with Discrete $X$**

$\text{educ}$: Respondent’s education:

1. 8 grades or less and no diploma or
2. 9-11 grades
3. High school diploma or equivalency test
4. More than 12 years of schooling, no higher degree
5. Junior or community college level degree (AA degrees)
6. BA level degrees; 17+ years, no postgraduate degree
7. Advanced degree
Nonparametric Regression with Discrete $X$

income: Respondent's family income:

- 1. None or less than $2,999
- 2. $3,000-$4,999
- 3. $5,000-$6,999
- 4. $7,000-$8,999
- 5. $9,000-$9,999
- 6. $10,000-$10,999
- ...
- 17. $35,000-$39,999
- 18. $40,000-$44,999
- ...
- 23. $90,000-$104,999
- 24. $105,000 and over
Marginal Distribution of $Y$
Income and Education

The graph shows the relationship between income and education. The x-axis represents the jitter of education levels, while the y-axis represents income levels. The data points are scattered across the plot, indicating a correlation between the two variables.
Distribution of income given education $p(y|x)$
Nonparametric Regression with Discrete $X$

- Hard to decode what is going on in the histograms
- Let’s try to find a more parsimonious summary measure: $E[Y|X]$
- Here our $X$ variable education has a small number of levels (7) and there are a reasonable number of observations in each level
- In situations like this we can estimate $E[Y|X = x]$ as the sample mean of $Y$ at each level of $x \in X$ (just like the binary case)
Nonparametric Regression with Discrete $X$
Nonparametric Regression

- This approach makes minimal assumptions
- It works well as long as
  - $X$ is discrete
  - there are a small number values of $X$
  - a small number of $X$ variables
  - a lot of observations at each $X$ value
- This method does not impose any specific \textbf{functional form} on the relationship between $Y$ and $X$ (i.e. the shape of $E[Y|X]$)
- $\rightarrow$ It is called a \textbf{nonparametric regression}

- But what do we do when $X$ is continuous and has many values?
Nonparametric Regression with Continuous $X$

Consider the Chirot data:


- Peasant Rebellions in Romanian counties in 1907
- Peasants made up 80% of the population
- About 60% of them owned no land which was mostly concentrated among large landowners

- We’re interested in the relationship between:
  - $Y$: intensity of the peasant rebellion
  - $X$: inequality of land tenure

- Around 11,000 peasants were killed by Romanian military
Nonparametric Regression with Continuous $X$
Uniform Kernel Regression: Simple Local Averages

- One approach is to use a **moving local average** to estimate $E[Y|X]$
- Calculate the average of the observed $y$ points that have $x$ values in the interval $[x_0 - h, x_0 + h]$
- $h = \text{some positive number (called the bandwidth)}$
- **Uniform kernel**: every observation in the interval is equally weighted

This gives the **uniform kernel regression**:

$$
\hat{E}[Y|X = x_0] = \frac{\sum_{i=1}^{N} K_h((X_i - x_0)/h)Y_i}{\sum_{i=1}^{N} K_h((X_i - x_0)/h)}
$$

where $K_h(u) = \frac{1}{2} \mathbf{1}_{\{|u| \leq 1\}}$
Uniform Kernel Regression: Simple Local Averages

![Graph showing inequality vs. intensity with data points and a green line highlighting certain areas.]

Stewart (Princeton)
Uniform Kernel Regression: Simple Local Averages
Changing the Bandwidth

Regression as an Asymmetric Summary of Bivariate Data
Regression as a Model of Conditional Expectation
Linear Regression

Adam Glynn
Gov2000: Quantitative Methodology for Political Science I
Week 4: Testing/Regression
October 3/5, 2016
Changing the Bandwidth

Regression as an Asymmetric Summary of Bivariate Data

Regression as a Model of Conditional Expectation

Linear Regression

Stewart (Princeton)

Week 4: Testing/Regression

October 3/5, 2016 81 / 133
Kernel Regression: Weighted Local Averages

- Another approach is to construct weighted local averages.
- Data points that are closer to $x_0$ get more weight than points farther away.

1. Decide on a symmetric, non-negative kernel weight function $K_h$ (e.g., Epanechnikov).

2. Compute weighted average of the observed $y$ points that have $x$ values in the bandwidth interval $[x_0 - h, x_0 + h]$ e.g.

$$\hat{E}[Y | X = x_0] = \frac{\sum_{i=1}^{N} K_h((X_i - x_0)/h)Y_i}{\sum_{i=1}^{N} K_h((X_i - x_0)/h)} \quad \text{where} \quad K_h(u) = \frac{3}{4}(1 - u^2)1_{\{|u| \leq 1\}}$$
Kernel Regression: Weighted Local Averages

Regression as an Asymmetric Summary of Bivariate Data
Regression as a Model of Conditional Expectation
Linear Regression

Adam Glynn
Gov2000: Quantitative Methodology for Political Science I

Stewart (Princeton)
Week 4: Testing/Regression
October 3/5, 2016 83 / 133
Changing the Bandwidth

Regression as an Asymmetric Summary of Bivariate Data

Regression as a Model of Conditional Expectation

Linear Regression

Adam Glynn
Gov2000: Quantitative Methodology for Political Science I

Week 4: Testing/Regression
October 3/5, 2016
Bias-Variance Tradeoff

When choosing an estimator $\hat{E}[Y|X]$ for $E[Y|X]$, we face a bias-variance tradeoff.

Notice that we can chose models with various levels of flexibility:

- A very flexible estimator allows the shape of the function to vary (e.g. a kernel regression with a small bandwidth).

- A very inflexible estimator restricts the shape of the function to a particular form (e.g. a kernel regression with a very wide bandwidth).
Hypothetical True Distribution

- Let’s conduct a simulation experiment to actually see the tradeoff
- Suppose we have the following population distribution:
Hypothetical True Distribution

- Another way of representing the same population distribution:

- From this distribution we draw thousands of simulated data sets.
An Example of Simulated Data Set
Two Estimators

- For each simulated data, we apply two simple estimators of $E(Y|X)$:
  - Divide $X$ into 4 ranges and take the mean for each
  - Divide $X$ into 8 ranges and take the mean for each

- We then evaluate how well these estimators do in terms of bias and variance.
Simulated Distribution of Estimates: 4 Intervals

inequality

intensity

Stewart (Princeton)
Simulated Distribution of Estimates: 8 Intervals

Inequality intensity

Stewart (Princeton)

Week 4: Testing/Regression

October 3/5, 2016
Bias-Variance Tradeoff

- A less “flexible” estimator leads to more bias
- A more “flexible” estimator leads to more variance
- As the name suggests, this problem cannot be “fixed”
- If we have more data or fewer variables we can “afford” to use a more flexible estimator
1. Testing: Making Decisions
   - Hypothesis testing
   - Forming rejection regions
   - P-values

2. Review: Steps of Hypothesis Testing

3. The Significance of Significance

4. Preview: What is Regression

5. Fun With Salmon

6. Nonparametric Regression
   - Discrete X
   - Continuous X
   - Bias-Variance Tradeoff

7. Linear Regression
   - Combining Linear Regression with Nonparametric Regression
   - Least Squares

8. Interpreting Regression

9. Fun With Linearity
Parametric Approach: Linear Regression

Linear regression works by assuming linear parametric form for the conditional expectation function:

\[ E[Y|X] = \beta_0 + X \beta_1 \]

- Conditional expectation defined by only two **coefficients** which are estimated from the data:
  - \( \beta_0 \) is called the **intercept** or **constant**
  - \( \beta_1 \) is called the **slope coefficient**

- Notice that the linear functional form imposes a constant slope
- Assumption: Change in \( E[Y|X] \) is the same at all values of \( X \)

- Geometrically, the linear regression function will look like:
  - A **line** in cases with a single \( X \) variable
  - A **plane** in cases with two \( X \) variables
  - A **hyperplane** in cases with more than two \( X \) variables
Parametric Approach: Linear Regression

Regression Line: intensity = −3.2 + inequality 5.1
Parametric Approach: Linear Regression

Warning: the model won’t always be a good fit for the data (even though it really wants to be)

Figure: ‘If I fits, I sits’

Linear regression always returns a line regardless of the data.
Interpretation of the regression slope

When we model the regression function as a line, we can interpret the parameters of the line in appealing ways:

1. **Intercept**: the average outcome among units with \( X = 0 \) is \( \beta_0 \):

   \[
   E[Y|X = 0] = r(0) = \beta_0 + \beta_1 0 = \beta_0
   \]

2. **Slope**: a one-unit change in \( X \) is associated with a \( \beta_1 \) change in \( Y \)

   \[
   E[Y|X = x + 1] - E[Y|X = x] = r(x + 1) - r(x) \\
   = (\beta_0 + \beta_1(x + 1)) - (\beta_0 + \beta_1 x) \\
   = \beta_0 + \beta_1 x + \beta_1 - \beta_0 - \beta_1 x \\
   = \beta_1
   \]
Linear regression with a binary covariate

Using the two facts above, it’s easy to see that when $X$ is binary, then we have the following:

1. **Intercept**: $E[Y|X = 0] = \beta_0$

2. **Slope**: average difference between $X = 1$ group and $X = 0$ group:
   $$\beta_1 = E[Y|X = 1] - E[Y|X = 0]$$

Thus, we can read off the difference in means between two groups as the slope coefficient on a linear regression.
Linear CEF with a binary covariate
Testing: Making Decisions
- Hypothesis testing
- Forming rejection regions
- P-values

Review: Steps of Hypothesis Testing

The Significance of Significance

Preview: What is Regression

Fun With Salmon

Nonparametric Regression
- Discrete X
- Continuous X
- Bias-Variance Tradeoff

Linear Regression
- Combining Linear Regression with Nonparametric Regression
- Least Squares

Interpreting Regression

Fun With Linearity
LOESS

- We can combine the nonparametric kernel method idea of using only local data with a parametric model
- Idea: fit a linear regression within each band
- Locally weighted scatterplot smoothing (LOWESS or LOESS):
  1. Pick a subset of the data that falls in the interval \([x - h, x + h]\)
  2. Fit a line to this subset of the data (\(\text{local linear regression}\)), weighting the points by their distance to \(x\) using a kernel function
  3. Use the fitted regression line to predict the expected value of \(E[Y|X = x_0]\)
Weighted Local Linear Regressions

Regression as an Asymmetric Summary of Bivariate Data
Regression as a Model of Conditional Expectation
Linear Regression

Adam Glynn
Gov2000: Quantitative Methodology for Political Science I

Week 4: Testing/Regression
October 3/5, 2016
Weighted Local Linear Regressions

Regression as an Asymmetric Summary of Bivariate Data
Regression as a Model of Conditional Expectation
Linear Regression

Testing: Making Decisions
- Hypothesis testing
- Forming rejection regions
- P-values

Review: Steps of Hypothesis Testing

The Significance of Significance

Preview: What is Regression

Fun With Salmon

Nonparametric Regression
- Discrete X
- Continuous X
- Bias-Variance Tradeoff

Linear Regression
- Combining Linear Regression with Nonparametric Regression
- Least Squares

Interpreting Regression

Fun With Linearity
Back up and review

To review our approach:

- We wanted to estimate the CEF/regression function $r(x) = E[Y|X = x]$, but it may be too hard to do nonparametrically.
- So we can model it: place restrictions on its functional form.
- Easiest functional form is a line:

$$r(x) = \beta_0 + \beta_1 x$$

- $\beta_0$ and $\beta_1$ are population parameters just like $\mu$ or $\sigma^2$!
- Need to estimate them in our samples! But how?
Simple linear regression model

- Let’s write our model as:

\[ Y_i = r(X_i) + u_i \]
\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]

- Now, suppose we have some estimates of the slope, \( \hat{\beta}_1 \), and the intercept, \( \hat{\beta}_0 \). Then the fitted or sample regression line is

\[ \hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \]
Fitted linear CEF/regression function

\[ \hat{r}(x) \]

Log Settler Mortality vs. Log GDP per capita growth.
Fitted linear CEF/regression function

\[ \hat{r}(x) \]
\[ \hat{r}_{alt}(x) \]

Log GDP per capita growth
Log Settler Mortality
Fitted values and residuals

**Definition (Fitted Value)**

A **fitted value** or **predicted value** is the estimated conditional mean of $Y_i$ for a particular observation with independent variable $X_i$:

$$\hat{Y}_i = \tilde{r}(X_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

**Definition (Residual)**

The **residual** is the difference between the actual value of $Y_i$ and the predicted value, $\hat{Y}_i$:

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$
Fitted linear CEF/regression function
Fitted linear CEF/regression function

Log Settler Mortality vs. Log GDP per capita growth

$\hat{r}(x)$

KOR

$Y_i$

$\hat{Y}_i$
Fitted linear CEF/regression function

$\hat{u}_i = Y_i - \hat{Y}_i$

$\hat{r}(x)$
Log Settler Mortality
Log GDP per capita growth

\[ \hat{r}_{\text{alt}}(x) \]

\[ \hat{u}_i = Y_i - \hat{Y}_i \approx 0 \]
Minimize the residuals

The residuals, $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$, tell us how well the line fits the data.

- Larger magnitude residuals means that points are very far from the line
- Residuals close to 0 mean points very close to the line

The smaller the magnitude of the residuals, the better we are doing at predicting $Y$

Choose the line that minimizes the residuals
Which is better at minimizing residuals?

\[ r(x) \quad r_{alt}(x) \]

Log GDP per capita growth
Log Settler Mortality
Minimizing the residuals

- Let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be possible values of the intercept and slope

- **Least absolute deviations (LAD) regression:**

$$
(\hat{\beta}_{0LAD}^{LAD}, \hat{\beta}_{1LAD}^{LAD}) = \arg \min_{\tilde{\beta}_0, \tilde{\beta}_1} \sum_{i=1}^{n} |Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_i|
$$

- **Least squares (LS) regression:**

$$
(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{\tilde{\beta}_0, \tilde{\beta}_1} \sum_{i=1}^{n} (Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_i)^2
$$

- Sometimes called **ordinary least squares (OLS)**
Why least squares?

- Easy to derive the least squares estimator
- Easy to investigate the properties of the least squares estimator
- Least squares is optimal in a certain sense that we’ll see in the coming weeks
Linear Regression: Justification

- Linear regression imposes a strong assumption on $E[Y|X]$
- Why would we ever want to do this?
  - Theoretical reason to assume linearity
  - Ease of interpretation
  - Bias-variance tradeoff
  - Analytical derivation of sampling distributions (next few weeks)
  - We can make the model more flexible, even in a linear framework (e.g. we can add polynomials, use log transformations, etc.)
Testing: Making Decisions
- Hypothesis testing
- Forming rejection regions
- P-values

Review: Steps of Hypothesis Testing

The Significance of Significance

Preview: What is Regression

Fun With Salmon

Nonparametric Regression
- Discrete X
- Continuous X
- Bias-Variance Tradeoff

Linear Regression
- Combining Linear Regression with Nonparametric Regression
- Least Squares

Interpreting Regression

Fun With Linearity
Linear Regression as a Predictive Model

- Linear regression can also be used to predict new observations
- Basic idea:
  - Find estimates $\hat{\beta}_0, \hat{\beta}_1$ of $\beta_0, \beta_1$ based on the in-sample data
  - To find the expected value of $Y$ for an out-of-sample data point with $X = x_{new}$ calculate:

$$\hat{E}[Y|X = x_{new}] = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$$

- While the line is defined over all regions of the data we may be concerned about:
  - interpolation
  - extrapolation
  - predicting in ranges of $X$ with sparse data
Which Predictions Do You Trust?

![Graphs showing scatter plots with regression lines.]

- The graphs illustrate the relationship between two variables, X and Y, with scattered data points.
- The regression lines are drawn to show the trend in the data.
- The graphs are used to analyze and trust predictions based on the observed data patterns.
In a 2004 *Nature* article, Tatem et al. use linear regression to conclude that in the year 2156 the winner of the women’s Olympic 100 meter sprint may likely have a faster time than the winner of the men’s Olympic 100 meter sprint.

How do the authors make this conclusion?

Using data from 1900 to 2004, they fit linear regression models of the winning 100 meter time on year for both men and women. They then use the estimates from these models to extrapolate 152 years into the future.
Tatem et al.'s predictions. Men's times are in blue, women's times are in red.
Alternate Models Fit Well, Yield Different Predictions
Alternate Models Fit Well, Yield Different Predictions

![Graph showing winning times over years]
The Trouble with Extrapolation

- The model only gives the best fitting line where we have data, it says little about the shape where there isn’t any data.
- We can always ask illogical questions and the model gives answers.
  - For example, when will women finish the sprint in negative time?
- Fundamentally we are assuming that data outside the sample looks like data inside the sample, and the further away it is the less likely that is to hold.
- Next semester we will talk about how this problem gets much harder in high dimensions
A More Subtle Example

Justices Become More Liberal With Time
Ideological ratings of Supreme Court justices as they age, since 1937

SOURCE: MARTIN-QUINN SCORES
A More Subtle Example

Nate Silver @NateSilver538 · Oct 5
So, basically, John Roberts is going to be Ruth Bader Ginsburg by 2036.
53eig.ht/1Gsl2u6

Supreme Court Justices Get More Liberal As They Get Older
The Supreme Court justices are back from vacation. They’ve picked up their robes from the cleaners — Alito’s had a pesky mustard stain — and are ...
Regression as Description/Prediction

- Even for simple problems regression can be challenging
- Always think about where we have data and what we are using to build our claims
- Summary: ‘prediction is hard, especially about the future’
Regression as a Causal Model (A Preview)

- Can regression be also used for causal inference?
- Answer: A very qualified yes

- For example, can we say that a one unit increase in inequality causes a 5.2 point increase in intensity?

- To interpret $\beta$ as a causal effect of $X$ on $Y$, we need very specific and often unrealistic assumptions:
  1. $E[Y|X]$ is correctly specified as a linear function (linearity)
  2. There are no other variables that affect both $X$ and $Y$ (exogeneity)

  (1) can be relaxed by:
  - Using a flexible nonlinear or nonparametric method
  - “Preprocessing” data to make analysis robust to misspecification

  (2) can be made plausible by:
  - Including carefully-selected control variables in the model
  - Choosing a clever research design to rule out confounding

- We will return to this later in the course

- For now, it is safest to treat $\beta$ as a purely descriptive/predictive quantity
Summary of Today

- Regression is about **conditioning**
- Regression can be used for **description, causation and prediction**
- Linear regression is a **parametrically restricted** form of regression
Next Week

- Basic linear regression
- Properties of OLS
- Reading:
  - Fox Chapter 6.1
  - Optional: Imai 4.2
Iterated learning: Intergenerational knowledge transmission reveals inductive biases

MICHAEL L. KALISH
University of Louisiana, Lafayette, Louisiana

THOMAS L. GRIFFITHS
University of California, Berkeley, California

AND

STEPHAN LEWANDOWSKY
University of Western Australia, Perth, Australia

Cultural transmission of information plays a central role in shaping human knowledge. Some of the most complex knowledge that people acquire, such as languages or cultural norms, can only be learned from other people, who themselves learned from previous generations. The prevalence of this process of “iterated learning” as a mode of cultural transmission raises the question of how it affects the information being transmitted. Analyses of iterated learning utilizing the assumption that the learners are Bayesian agents predict that this process should converge to an equilibrium that reflects the inductive biases of the learners. An experiment in iterated function learning with human participants confirmed this prediction, providing insight into the consequences of intergenerational knowledge transmission and a method for discovering the inductive biases that guide human inferences.
The Design

- Each learner sees a set of \((x, y)\) pairs
- Makes predictions of \(y\) for new \(x\) values
- Predictions are data for the next learner
Results
References
