Week 4: Testing/Regression

Brandon Stewart

Princeton

September 21–25, 2020

---

These slides are heavily influenced by Matt Blackwell, Adam Glynn, Jens Hainmueller, and Erin Hartman
Where We’ve Been and Where We’re Going...

- Last Week
  - inference and estimator properties
  - point estimates, confidence intervals
- This Week
  - hypothesis testing
  - what is regression?
  - nonparametric and linear regression
- Next Week
  - inference for simple regression
  - properties of ordinary least squares
- Long Run
  - probability → inference → regression → causal inference
1 Hypothesis Testing
   - Terminology and Procedure
   - One-Sided Tests
   - Connections
   - Power

2 p-values
   - Mechanics
   - Multiple Testing
   - Fun With Salmon
   - The Significance of Significance

3 What is Regression?
   - Conditional Expectation Functions
   - Nonparametric Regression
   - Best Linear Predictor
   - Ordinary Least Squares

4 Interpreting Regression
   - Fun With Linearity
Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment

ALAN S. GERBER  Yale University
DONALD P. GREEN  Yale University
CHRISTOPHER W. LARIMER  University of Northern Iowa
We Secretly Already Covered This!

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We’ve been talking about the problem for years, but it only seems to get worse. This year, we’re taking a new approach. We’re sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

<table>
<thead>
<tr>
<th>MAPLE DR</th>
<th>Aug 04</th>
<th>Nov 04</th>
<th>Aug 06</th>
</tr>
</thead>
<tbody>
<tr>
<td>9995 JOSEPH JAMES SMITH</td>
<td>Voted</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9995 JENNIFER KAY SMITH</td>
<td>Voted</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9997 RICHARD B JACKSON</td>
<td>Voted</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9999 KATHY MARIE JACKSON</td>
<td>Voted</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Our estimand is the difference in population means: \( \theta = \mu_y - \mu_x \).
Where \( \mu_y \) is those receiving the social pressure mailer and \( \mu_x \) is those receiving the civic duty mailer.

- We used difference in means to get an estimate of 0.063.
- We estimated the estimator’s standard error \( \hat{SE}(\hat{\theta}) = 0.02 \).
- What if there was no difference in means in the population \( (\mu_y - \mu_x = 0) \)?
- By asymptotic Normality \( (\hat{\theta} - 0)/SE(\hat{\theta}) \sim N(0, 1) \)
- This implies that \( \hat{\theta} \sim N(0, SE(\hat{\theta})) \)
We Secretly Already Covered This!

Example from Gerber, Green and Larimer (2008).
We assumed we knew the difference in means was zero.

We derived the sampling distribution under that value.

We asked ‘if the population difference was zero, how likely would it be to see an observe difference as extreme (or more extreme) than our observed estimate?’

Our observed difference was so implausible we concluded it was unlikely the population difference was really zero.
What is a Hypothesis Test?

- A **hypothesis test** is a way of assessing evidence against a particular hypothesis about the world—it is a statistical **thought experiment**.
- By using probability, we define what the data **would look like** under a given state of the world, and then assess whether or not that’s consistent with our **observed data**.
- Hypothesis testing is an alternative way to think about **inference** than confidence intervals, but using much of the same infrastructure.
- Hypothesis tests lead to **discrete** decisions.
An Example from Drug 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.
### Example

<table>
<thead>
<tr>
<th>Subject</th>
<th>SBP(_{before})</th>
<th>SBP(_{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>
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>

Two kinds of bad decisions:

- False positive
- False negative
Hypothesis Testing

- Terminology and Procedure
  - One-Sided Tests
  - Connections
  - Power

p-values

- Mechanics
- Multiple Testing
- Fun With Salmon
- The Significance of Significance

What is Regression?

- Conditional Expectation Functions
- Nonparametric Regression
- Best Linear Predictor
- Ordinary Least Squares

Interpreting Regression

- Fun With Linearity
Elements of a Hypothesis Test

Important terms we are about to define:

- Null Hypothesis (assumed state of world for test)
- Alternative Hypothesis (all other states of the world)
- Type I and Type II Errors (two types of errors)
- Test Statistic (what we will observe from the sample)
- Test Level (the probability of a type I error)
- Rejection Region (the basis of our decision)
Hypotheses

- **Null Hypothesis** ($H_0$): The conservatively assumed state of the world we are accumulating evidence against (often “no effect”).

  Example: There is not a difference in voting rates between those who received the social pressure mailer and those that received the civic duty mailer.

- **Alternative Hypothesis** ($H_a$): The state of the world where the null hypothesis is not true and thus the claim to be indirectly tested.

  Example: There is a difference in voting rates between those who received the social pressure mailer and those that received the civic duty mailer.
## Error Types

<table>
<thead>
<tr>
<th></th>
<th>((H_0 \text{ False}))</th>
<th>((H_0 \text{ True}))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reject (H_0)</strong></td>
<td>Correct</td>
<td>Type I error</td>
</tr>
<tr>
<td><strong>Don’t Reject (H_0)</strong></td>
<td>Type II error</td>
<td>Correct</td>
</tr>
</tbody>
</table>

We generally make the normative judgment that we prefer an undetected finding (Type II error) to a false discovery (Type I error).
A Visual Reminder from Allison Horst

Type 1 Errors

[p = 0.02]

(sigh...) Yes.
I'm still sure
I birthed BOTH of you.

OK, what about NOW??

Artwork by @allison_horst
A Visual Reminder from Allison Horst

**Type II Errors:**

That's not how you can't just... I mean, no matter how close you get... ugh.

\[ p = 0.31 \]

Artwork by @allison_horst
**Test Statistic**: which we denote $T_n$ is a function of the sample, the estimator and the null hypothesis.

$$T_n = \frac{\hat{\theta} - \theta_0}{\text{SE}[\hat{\theta}]}$$

This is a random variable because it is a function of the sample.

**Null Distribution**: the sampling distribution of the statistic/test statistic assuming that the null is true.
Test Statistics and Null Distribution

Returning to the voting experiment. We know from the Central Limit Theorem that the standardized difference in means has a standard normal distribution asymptotically,

\[ T_n = \frac{\hat{\theta} - (\mu_y - \mu_x)}{\text{SE}[\hat{\theta}]} \xrightarrow{d} \mathcal{N}(0, 1) \]

Under the null hypothesis of \( \mu_y - \mu_x = 0 \), we have

\[ T_n = \frac{\hat{\theta}}{\text{SE}[\hat{\theta}]} \xrightarrow{d} \mathcal{N}(0, 1) \]

If \( T_n \) is very far from zero—in the sense that it has low probability under \( \mathcal{N}(0, 1) \)—then we reject the null hypothesis as not plausible.
Rejection Region

- The rejection region $R$ contains the values of the test statistics, $T_n$ for which we reject the null. This is defined by the null distribution and our chosen tolerance for Type I error.
- Denote the probability of Type I error as $\alpha$.
- If we have a two-sided test (i.e. our null hypothesis equals a given value and thus extreme values on either side could reject the null), we reject when $|T_n| > c$ where $P_{H_0}(|T_n| > c) = \alpha$.
- We call $c$ the critical value.
- Much like the 95% confidence interval, we pick $\alpha = .05$ by convention.
The value for which $P=0.05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty.

- Ronald Fisher, *Design of Experiments* (1922)
Two-sided Rejection Region

Rejection region with $\alpha = 0.05$, $H_0 : \theta = 0$, $H_A : \theta \neq 0$.
We define the rejection region such that $|T_n| > c$ where $P_{H_0}(|T_n| > c) = \alpha$.

This means that we will get a false rejection only 5% of the time.

We want to find the point $c$ such that $P_{H_0}(T_n < c) + P_{H_0}(T_n > c) = \alpha$ where we typically use equal probability on each side by convention.

This is just the task of finding the quantile for $\alpha/2$. In the case of $\alpha = .05$, $qnorm(.05/2) = 1.96$
The Complete Recipe

1. Define a null hypothesis and the alternative hypothesis.
2. Choose a test statistic $T_n$.
3. Determine the Type I error you will tolerate ($\alpha$).
4. Determine your critical value and thus your rejection region.
5. Calculate your test statistic in your observed data.
6. If your observed data is sufficiently unlikely under your null hypothesis, reject your null.
diff <- mean(treated) - mean(control)
se_diff <- sqrt(var(treated)/length(treated) + var(control)/length(control))
test_statistic <- diff/se_diff

This yields 18.3 which is much better than our .05 critical value of 1.96.

We reject the null.
Let’s go back to the FDA making a decision about drugs.

<table>
<thead>
<tr>
<th>FDA approves</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>

Drug trials are expensive and *ex ante* we can specify that we only care about one direction in particular. Consider the Sowers et al (2006) case which claimed to decrease blood pressure.
We can define our hypotheses for a one-sided test.

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

We can calculate the test statistic:

- \( \bar{x} = 21.0 \)
- \( \hat{\sigma} = 14.3 \)
- \( n = 345 \)

Therefore,

\[ t_n = \frac{21.0 - 0}{\frac{14.3}{\sqrt{345}}} = 27.3 \]

We construct our rejection region with \( c = qnorm(.95) = 1.644 \).
Rejection Region with $\alpha = .05$
Rejection Region with $\alpha = .05$

Test Statistic $T = 27.3$
This strategy works for any asymptotically normal estimator.

A size-$\alpha$ $t$-test (also called a Wald test) rejects $H_0$ when $|T_n| > c$ where

$$T_n = \frac{\hat{\theta} - \theta_0}{\hat{SE}[\hat{\theta}]}$$

We get the critical value $c$ by using the standard normal to get the probability such that $P_{H_0}(T_n \leq c) = 1 - \alpha/2$

This will guarantee the nominal probability of Type I error as $n$ gets large.
Connections to Confidence Intervals

You may have noticed that there is a 1:1 mapping between CIs and hypothesis tests (same math, different story).

- If the confidence interval includes the null hypothesis, fail to reject your null.
- If the confidence interval does not include the null hypothesis, reject your null.

A $100(1 - \alpha)\%$ confidence interval represents all null hypotheses that we would not reject with a $\alpha$ level test.

We can think of confidence intervals as a range of plausible values in the sense that we would not have rejected them had they been our null hypotheses.
1. Hypothesis Testing
   - Terminology and Procedure
   - One-Sided Tests
   - Connections
   - Power

2. p-values
   - Mechanics
   - Multiple Testing
   - Fun With Salmon
   - The Significance of Significance

3. What is Regression?
   - Conditional Expectation Functions
   - Nonparametric Regression
   - Best Linear Predictor
   - Ordinary Least Squares

4. Interpreting Regression
   - Fun With Linearity
Power

We designed our tests to minimize Type I error (the false positive). However we might also worry we are failing to detect important findings. Experiments and surveys are expensive and so we want to design our work to minimize the chance that we fail to reject the null at the end of the day.

**Definition (Power)**

The power of a test is the probability that a test rejects the null given some assumed population distribution $P_{\theta}(|T_n| > c)$.

$$\text{Power} = 1 - P(\text{Type II error})$$

If we fail to reject the null hypothesis, there are basically three possible states of the world:

1. Null is true (no difference between the mailer populations).
2. Null is false (there is a difference between the mailer populations), but test had low power.
3. Null is false, the test is well-powered and we got incredibly unlucky.
Imagine that you are studying whether there are differences in college admissions by race.

You take a sample of 20 applications and conduct a hypothesis test of the difference in admissions rates.

You fail to reject the null. Is this good evidence that there are no differences?

No! You haven’t been able to confidently reject the possibility that there is no difference, but that’s different than rejecting the possibility that there is a difference.

This might seem obvious but conflating a lack of evidence with evidence for a zero effect is a problem that crops up in a lot of work.

Power analysis is a way of guiding the choice of sample size prior to an experiment to avoid this kind of mistake.
Steps for Power Analysis

1) Specify the null hypothesis to be tested at significance level $\alpha$

2) Choose a true value for population parameters and derive the sampling distribution of test statistic

3) Calculate the probability of rejecting the null hypothesis under this sampling distribution.

4) Find the smallest sample size such that this rejection probability equals a pre-specified power level.

5) Possibly repeat under different assumptions about the population.
Why did they use 38 thousand people per group? Power analysis!
If their results were exactly true, would we be expect to be able to confirm that the effect is non-zero with a replication using only 500 mailers?
For this example, let’s ignore the household sampling, but see the challenge problem from Course Meeting 3.
### What Do We Know?

\[
\begin{align*}
\mu_y &= .378 \\
\sigma^2_y &= (.378)(1 - .378) = 0.235 \\
\mu_x &= .315 \\
\sigma^2_x &= (.315)(1 - .315) = 0.216 \\
\theta &= \mu_y - \mu_x = .063 \\
\theta_0 &= 0 \\
V[\hat{\theta}_n] &= .235/250 + 0.216/250 = .001804 \\
\hat{\theta}_n &\xrightarrow{d} \mathcal{N}(0.63, 0.001804)
\end{align*}
\]

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Civic Duty</th>
<th>Hawthorne</th>
<th>Self</th>
<th>Neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage Voting</td>
<td>29.7%</td>
<td>31.5%</td>
<td>32.2%</td>
<td>34.5%</td>
<td>37.8%</td>
</tr>
<tr>
<td>N of Individuals</td>
<td>191,243</td>
<td>38,218</td>
<td>38,204</td>
<td>38,218</td>
<td>38,201</td>
</tr>
</tbody>
</table>
Calculating Probability of Rejecting the Null

- We reject when
  
  \[ |T| = \frac{|\hat{\theta}_n - 0|}{\text{SE}[\hat{\theta}_n]} > 1.96 \]

- Rearranging we get rejection when
  
  \[ |\hat{\theta}_n| > 1.96\text{SE}[\hat{\theta}_n] \]

- Under our assumption of the truth we know \( V[\hat{\theta}_n] = .001804 \) and thus:
  
  \[ \{ \hat{\theta}_n < -1.96\sqrt{.001804} \} \cup \{ \hat{\theta}_n > 1.96\sqrt{.001804} \} \]

- Using the sampling distribution we derived can calculate:
  
  \[ P\left( \hat{\theta}_n < -1.96\sqrt{.001804} \right) + P\left( \hat{\theta}_n > 1.96\sqrt{.001804} \right) \]
Getting an Answer in R

Let’s calculate this using the fact that $\hat{\theta}_n \overset{d}{\to} N(0.63, 0.001804)$

```
se <- sqrt(.001804)
pnorm(-1.96*se, mean=.05, sd=se) +
pnorm(1.96*se, mean=.05, sd=se, lower.tail=FALSE)
```

We get 0.2177 which tells us that if the true population means were those calculated in the experiment, we would be able to reject the null of no effect about 22% of the time using 500 mailers.

Yikes! That is not well powered.
True Difference = 0.05, Power = 0.22
Wait! Does that mean I need to know the truth before I start to do power analysis?

Yeah...kind of. In practice, this is why we calculate under many possible configurations.
Power Curve

You can graph power for various possible effect sizes (here for 500, 1000, and 10000 samples).
Power calculations

- Power calculations depend on your assumed difference and the population variance, neither of which is typically known.
- Instead you may want to power depending on the minimal difference of interest with a conservative (high) estimate of population variance.
- In general power is favorable when:
  - large \( n \)
  - bigger difference (pushes the alternative distribution away)
  - smaller variance (it squeezes the distribution in)
- Power analysis is really important if you are planning experiments, but we will touch on it only cursorily in this class. The Gerber and Green Field Experiments book is an amazing resource for more on experiments in general.
We Covered

- 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.
- There is a close relationship between the results of an $\alpha$ level hypothesis test and the coverage of a $(1 - \alpha)\%$ confidence interval.
- Power analysis is a way to assess your probability of missing a finding of interest.
- We will cover more on subtleties of interpretation in future videos.

Next time: $p$-values!
Where We’ve Been and Where We’re Going...

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

- **This Week**
  - hypothesis testing
  - what is regression?
  - nonparametric and linear regression

- **Next Week**
  - inference for simple regression
  - properties of ordinary least squares

- **Long Run**
  - probability → inference → regression → causal inference
1 Hypothesis Testing
   - Terminology and Procedure
   - One-Sided Tests
   - Connections
   - Power

2 p-values
   - Mechanics
   - Multiple Testing
   - Fun With Salmon
   - The Significance of Significance

3 What is Regression?
   - Conditional Expectation Functions
   - Nonparametric Regression
   - Best Linear Predictor
   - Ordinary Least Squares

4 Interpreting Regression
   - Fun With Linearity
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.

**Definition ($p$-value)**

The $p$-value is the smallest value $\alpha$ such that an $\alpha$-level test would reject the null hypothesis.

Under the null hypothesis, this corresponds to the probability of observing a test statistic as extreme or more extreme than the one in the observed data (where extreme is defined in terms of the alternative hypothesis).
**p-values**

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

- \( H_0 : \theta \leq 0 \)  
  \( H_a : \theta > 0 \)

- \( H_0 : \theta = 0 \)  
  \( H_a : \theta \neq 0 \)

- \( H_0 : \theta \geq 0 \)  
  \( H_a : \theta < 0 \)

\[ T = 1.8 \]

\[ p = 0.036 \]

\[ p = .072 \]

\[ p = 0.964 \]
Rejection Regions and $p$-values

- Recall for the social pressure experiment we got a test statistics of $t_{obs} = 18.5$.
- How likely was it we would get a test statistics this extreme or more extreme under our null hypothesis?

$$P_0(|T_n| > 18.5) = P_0(T_n > 18.5) + P_0(T_n < -18.5) = 2P_0(T_n < -18.5)$$

- We can get this in R with $2 \times \text{pnorm}(-18.5)$
- That yields a $p$-value of $2.06 \times 10^{-76}$.
- By convention we would say it is statistically significant at level $\alpha$ for some $\alpha$ that the $p$-value is below.
The Sowers et. al. Example

Test Statistic

\[ p(T) \]

Fail to Reject

Reject

\[ p < .0001 \]

\[ T = 27.3 \]
All those guarantees on Type I error?

It only works for one test.
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:
## Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | -0.0280393 | 0.1138198 | -0.246  | 0.80605 |
| X2       | -0.1503904 | 0.1121808 | -1.341  | 0.18389 |
| X3       | 0.0791578  | 0.0950278 | 0.833   | 0.40736 |
| X4       | -0.0717419 | 0.1045788 | -0.686  | 0.49472 |
| X5       | 0.1720783  | 0.1140017 | 1.509   | 0.13518 |
| X6       | 0.0808522  | 0.1083414 | 0.746   | 0.45772 |
| X7       | 0.1029129  | 0.1141562 | 0.902   | 0.37006 |
| X8       | -0.3210531 | 0.1206727 | -2.661  | 0.00945 **|
| X9       | -0.0531223 | 0.1079834 | -0.492  | 0.62412 |
| X10      | 0.1801045  | 0.1264427 | 1.424   | 0.15827 |
| X11      | 0.1663864  | 0.1109471 | 1.500   | 0.13768 |
| X12      | 0.0080111  | 0.1037663 | 0.077   | 0.93866 |
| X13      | 0.0002117  | 0.1037845 | 0.002   | 0.99838 |
| X14      | -0.0659690 | 0.1122145 | -0.588  | 0.55829 |
| X15      | -0.1296539 | 0.115753  | -1.162  | 0.24872 |
| X16      | -0.0544456 | 0.1251395 | -0.435  | 0.66469 |
| X17      | 0.0043351  | 0.1120122 | 0.039   | 0.96923 |
| X18      | -0.0807963 | 0.1098525 | -0.735  | 0.46421 |
| X19      | -0.0858057 | 0.1185529 | -0.724  | 0.47134 |
| X20      | -0.1860057 | 0.1045602 | -1.779  | 0.07910 .|
| X21      | 0.0021111  | 0.1081179 | 0.020   | 0.98447 |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1 0

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: \( \frac{1}{20} = 0.05 \) of the tests are false positives at the 0.05 level.
- Also note that \( \frac{2}{20} = 0.1 \) are significant at the 0.1 level. Totally expected!
- The procedure by which tests/analyses are performed and shown to us matters a lot!
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.
This all seems rather bad. Fixing this is an active research area for a lot of people.

Two Styles of Solutions:
(1) statistical
and
(2) procedural.
Statistical Paths Forward

- Control the **Family-wise Error Rate**:
  
  
  \[ P(\text{making at least one Type I error}) \]

  - Bonferroni correction: reject the \( j \)th null hypothesis \( H_j \) if \( p_j < \alpha/m \) where \( m \) is the total number of tests
    
    ★ NB: VERY conservative

- Control the **False Discovery Rate**

  \[ E\left[ \frac{\# \text{ of false rejections}}{\max(\text{total \# of rejections}, 1)} \right] \]

  - Benjamini-Hochberg Procedure: Order the \( p \)-values
    
    \( p(1) \leq p(2) \leq \cdots \leq p(m) \)
    
    ★ Find the largest \( k \) such that \( p(i) \leq \frac{\alpha \times k}{m} \) and call it \( k^* \)
    
    ★ Reject all \( H_k \) for \( k \leq k^* \)
Benjamini-Hochberg Example

Pvalues: 0.011, 0.13, 0.06, 0.54, 0.008, 0.024, 0.001, 0.201, 0.78, 0.023

<table>
<thead>
<tr>
<th>Step 1:</th>
<th>Step 2:</th>
<th>Step 3:</th>
<th>Step 4:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sort</td>
<td>Find New Threshold</td>
<td>Find $k$</td>
<td>Reject</td>
</tr>
<tr>
<td>0.001</td>
<td>0.05*1/10 = 0.005</td>
<td>*</td>
<td>Reject</td>
</tr>
<tr>
<td>0.008</td>
<td>0.05*2/10 = 0.01</td>
<td>*</td>
<td>Reject</td>
</tr>
<tr>
<td>0.011</td>
<td>0.05*3/10 = 0.015</td>
<td>*</td>
<td>Reject</td>
</tr>
<tr>
<td>0.023</td>
<td>0.05*4/10 = 0.02</td>
<td></td>
<td>Reject</td>
</tr>
<tr>
<td>0.024</td>
<td>0.05*5/10 = 0.025</td>
<td>*</td>
<td>Reject</td>
</tr>
<tr>
<td>0.06</td>
<td>0.05*6/10 = 0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.13</td>
<td>0.05*7/10 = 0.035</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.201</td>
<td>0.05*8/10 = 0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.54</td>
<td>0.05*9/10 = 0.045</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.78</td>
<td>0.05*10/10 = 0.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Procedural Paths Forward

- **Preregistration**
  - in theory, forces people to pre-commit to analyses.
  - doesn’t directly address multiple comparisons, but may credibly limit them.
  - doesn’t work if pre-analysis plans aren’t abided by or if people register many, many hypotheses.

- **Sample Splits**
  - Set-aside one sample for discovery where you can search over lots of different options.
  - A second sample can be used to test a small set of hypotheses.
  - Similar to preregistration in that it doesn’t directly address multiple comparisons, but limits them.
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 mm$^3$ with a cluster-level significance of $p = .001$.\"
Okay, but what do they mean?
The Meaning of $p$-values (courtesy of XKCD)

<table>
<thead>
<tr>
<th>$p$-value</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>Highly significant</td>
</tr>
<tr>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>0.02</td>
<td>Highly suggestive</td>
</tr>
<tr>
<td>0.03</td>
<td>On the edge of significance</td>
</tr>
<tr>
<td>0.04</td>
<td>Oh crap. Redo calculations.</td>
</tr>
<tr>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>Significant at the $p &lt; 0.10$ level</td>
</tr>
<tr>
<td>0.051</td>
<td></td>
</tr>
<tr>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>0.07</td>
<td>Hey, look at this interesting subgroup analysis</td>
</tr>
<tr>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>$\geq 0.1$</td>
<td></td>
</tr>
</tbody>
</table>
The value of the $p$-value

Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.

*Ronald Fisher (1935)*

In social science (and I think in psychology as well), the null hypothesis is almost certainly false, false, false, and you don’t need a $p$-value to tell you this. The $p$-value tells you the extent to which a certain aspect of your data are consistent with the null hypothesis. A lack of rejection doesn’t tell you that the null hypothesis is likely true; rather, it tells you that you don’t have enough data to reject the null hypothesis.

*Andrew Gelman (2010)*
Practical versus Statistical Significance

\[ T_n = \frac{\hat{\theta} - \theta_0}{SE[\hat{\theta}]} = \frac{\bar{X} - \mu_0}{\sqrt{V[X]/n}} \]

- What are the possible reasons for rejecting the null?
  1. \( \bar{X} - \mu_0 \) 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. \( V[X] \) 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!
Problems with \( p \)-values

- \( p \)-values are extremely **common** in the social sciences and are often the standard by which the value of the finding is judged.
- \( p \)-values do **not**:
  - measure the size or importance of a result.
  - measure the probability that the alternative hypothesis is false.
  - measure the probability that the null hypothesis is true.
  - indicate how predictive a variable is of another.
  - provide a good indication of whether a paper should be published.
- A large \( p \)-value could mean either that we are in the null world OR that we had insufficient power.

See also: The ASA’s (American Statistical Association) Statement on \( p \)-Values: Context, Process, and Purpose
(http://dx.doi.org/10.1080/00031305.2016.1154108)
Arbitrary Publication Cutoffs

Figure 1a: Histogram of Z-Statistics, APSR & AJPS (Two-Tailed)

Gerber and Malhotra (2006) Top Political Science Journals
Arbitrary Publication Cutoffs

Figure 1
Histogram of $z$ Statistics From the *American Sociological Review*, the *American Journal of Sociology*, and *The Sociological Quarterly* (Two-Tailed)

Gerber and Malhotra (2008) Top Sociology Journals
Arbitrary Publication Cutoffs

Figure 1. The graphs show the distribution of 3,627 p values from three major psychology journals.

Masicampo and Lalande (2012) Top Psychology Journals
Viewpoint

The harm done by tests of significance

Ezra Hauer*

35 Merton Street, Apt. 1706, Toronto, Ont., Canada M4S 3G4

Abstract

Three historical episodes in which the application of null hypothesis significance testing (NHST) led to the mis-interpretation of data are described. It is argued that the pervasive use of this statistical ritual impedes the accumulation of knowledge and is unfit for use.

© 2003 Elsevier Ltd. All rights reserved.

Keywords: Significance; Statistical hypothesis; Scientific method
Table 1
The Virginia RTOR study

<table>
<thead>
<tr>
<th></th>
<th>Before RTOR signing</th>
<th>After RTOR signing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fatal crashes</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Personal injury crashes</td>
<td>43</td>
<td>60</td>
</tr>
<tr>
<td>Persons injured</td>
<td>69</td>
<td>72</td>
</tr>
<tr>
<td>Property damage crashes</td>
<td>265</td>
<td>277</td>
</tr>
<tr>
<td>Property damage (US$)</td>
<td>161243</td>
<td>170807</td>
</tr>
<tr>
<td>Total crashes</td>
<td>308</td>
<td>337</td>
</tr>
</tbody>
</table>
The Point in Hauer

- Two other interesting examples in Hauer (2004)
- Core issue is that lack of significance is not an indication of a zero effect, it could also be a lack of power (i.e. a small sample size relative to the difficulty of detecting the effect)
- On the opposite end, large tech companies rarely use significance testing because they have huge samples which essentially always find some non-zero effect. But that doesn’t make the finding significant in a colloquial sense of important.
What if I need to show evidence of a zero effect?
An Equivalence Approach to Balance and Placebo Tests

Erin Hartman  University of California Los Angeles
F. Daniel Hidalgo  Massachusetts Institute of Technology

Abstract: Recent emphasis on credible causal designs has led to the expectation that scholars justify their research designs by testing the plausibility of their causal identification assumptions, often through balance and placebo tests. Yet current practice is to use statistical tests with an inappropriate null hypothesis of no difference, which can result in equating nonsignificant differences with significant homogeneity. Instead, we argue that researchers should begin with the initial hypothesis that the data are inconsistent with a valid research design, and provide sufficient statistical evidence in favor of a valid design. When tests are correctly specified so that difference is the null and equivalence is the alternative, the problems afflicting traditional tests are alleviated. We argue that equivalence tests are better able to incorporate substantive considerations about what constitutes good balance on covariates and placebo outcomes than traditional tests. We demonstrate these advantages with applications to natural experiments.

Replication Materials: The data, code, and any additional materials required to replicate all analyses in this article are available on the American Journal of Political Science Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/RYNSDG.
Solution: Flip the hypotheses:

\[ H_0 : \theta_T - \theta_C \leq \epsilon_L \text{ or } \theta_T - \theta_C \geq \epsilon_U \]

versus

\[ H_A : \epsilon_L < \theta_T - \theta_C < \epsilon_U \]
Tests of Difference vs. Equivalence Tests

\[ H_0 : \frac{\mu_T - \mu_C}{\sigma} = 0 \quad \text{versus} \quad H_A : \frac{\mu_T - \mu_C}{\sigma} \neq 0 \]

Type I Error

Test has \( \alpha \) probability of declaring the two means different when they are, in fact, the same.

Problem: Controlling for the incorrect type of error if we’re trying to provide evidence in favor of equivalence.
Tests of Difference vs. **Equivalence Tests**

\[ H_0 : \frac{\mu_T - \mu_C}{\sigma} \geq \epsilon_U \quad \text{or} \quad \frac{\mu_T - \mu_C}{\sigma} \leq \epsilon_L \quad \text{versus} \quad H_A : \epsilon_L < \frac{\mu_T - \mu_C}{\sigma} < \epsilon_U \]

**Type I Error**

Test has \( \alpha \) probability of declaring the two means equivalent when they are, in fact, the different.

**Solution:** Now control for the correct type of false positive.
General Message:

Don’t misinterpret, or rely too heavily, on your p-values. They are evidence against your null, not evidence in favor of your alternative.
But Let’s Not Obsess Too Much About $p$-values

From Leek and Peng (2015) “$P$ values are just the tip of the iceberg” *Nature.*
We Covered

- *p*-values
- multiple testing
- the problems with *p*-values

Next Time: What is Regression?

Bonus reading for those interested to learn more:

- Hauer. 2004 “The harm done by tests of significance.” *Accident Analysis & Prevention.*
- Nuzzo. 2014. “Statistical Errors.” *Nature*
- Cohen. 1994. “The Earth is Round (*p* < 0.05).” *American Psychologist*
Where We’ve Been and Where We’re Going...

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

- **This Week**
  - hypothesis testing
  - what is regression?
  - nonparametric and linear regression

- **Next Week**
  - inference for simple regression
  - properties of ordinary least squares

- **Long Run**
  - probability → inference → regression → causal inference
1 Hypothesis Testing
   - Terminology and Procedure
   - One-Sided Tests
   - Connections
   - Power

2 p-values
   - Mechanics
   - Multiple Testing
   - Fun With Salmon
   - The Significance of Significance

3 What is Regression?
   - Conditional Expectation Functions
   - Nonparametric Regression
   - Best Linear Predictor
   - Ordinary Least Squares

4 Interpreting Regression
   - Fun With Linearity
What You’ve Probably Probably Seen This

We are going to go about this a slightly different way.
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.

- Does turnout vary by types of mailers received?
- Does the probability of getting a job vary by applicant race?
- Is intergenerational mobility less likely for children with an incarcerated parent?

By convention variables tend to be called $X$ and $Y$.

$Y$ - the dependent variable or outcome or regressand or left-hand-side variable or response

- Voter turnout
- Receiving a job
- 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
- Applicant race
- Incarcerated parent
Characterizing the Joint Distribution

- A first step in this process is to characterize the joint distribution between the two random variables, \( f_{X,Y} \), based on pairs of draws for the same unit \((X_i, Y_i)\).

- Generally we are trying to characterize some properties of the conditional distribution \( f_{Y|X} \) and often we will use the conditional expectation function, \( \mu(x) = E[Y|X = x] \), as a summary.

- We can use the conditional expectation function to help us perform important social science tasks:
  - **Description**: what is average value of \( Y \) among people with \( X = x \) in the population.
  - **Prediction**: for a random sample of the population and given a value of \( x \), \( \mu(x) \) is the best predictor in terms of squared error.
  - **Causal Inference**: with additional assumptions (later in the semester) we can talk about how intervening to change the value of \( X \) will change \( Y \).
Reminder of Definitions

Definition (Conditional Expectation (Discrete))
Let $Y$ and $X$ be discrete random variables. The conditional expectation of $Y$ given $X = x$ is defined as:

$$E[Y|X = x] = \sum_y y \cdot P(Y = y|X = x) = \sum_y y \cdot p_{Y|X}(y|x)$$

Definition (Conditional Expectation (Continuous))
Let $Y$ and $X$ be continuous random variables. The conditional expectation of $Y$ given $X = x$ is given by:

$$E[Y|X = x] = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y|x) dy$$
Implications of the CEF Definition

Let $X$ and $Y$ be random variables and decompose $Y_i$ into the conditional expectation function plus error such that,

$$Y_i = E[Y_i|X_i] + \epsilon_i$$

- Then, the expectation of the error doesn’t depend on $X_i$
  $$E[\epsilon_i|X_i] = E[\epsilon_i] = 0$$

- the error is uncorrelated with any function of $X_i$
  $$\text{Cov}[g(X_i), \epsilon_i] = 0$$

- the variance of the outcome given $X$ is the variance of the error given $X$.
  $$V[\epsilon_i|X_i] = V[Y_i|X_i]$$

- and the CEF is the **lowest mean squared error** predictor of $Y_i$ given $X_i$
CEF for binary covariates

- We’ve been writing $\mu_1$ and $\mu_0$ for the means in different groups.
- For example, on the problem set, we looked at the expected value of the loan amount given income group.
- Note that these are just conditional expectations. Define $Y$ to be the loan amount, $X = 1$ to indicate the high income group and $X = 0$ to indicate the low income group:

$$
\mu_1 = E[Y|X = 1]
$$

$$
\mu_0 = E[Y|X = 0]
$$

- Notice here that since $X$ can only take on two values, 0 and 1, then these two conditional means completely summarize the CEF.
- Estimation just involves taking the means within the groups.
Non-binary CEFs

- If $X$ is discrete with not too many categories, we can estimate the conditional expectation using the means within groups:
  - $\hat{\mu}(1) = \hat{E}[Y|X = 1] = \frac{1}{n_{x=1}} \sum_{i: X_i = 1} Y_i$
  - $\hat{\mu}(2) = \hat{E}[Y|X = 2] = \frac{1}{n_{x=2}} \sum_{i: X_i = 2} Y_i$
  - $\ldots$

- This kind of estimation is nonparametric in the sense that it makes no assumptions about the specific functional form of $\mu(x)$.

- When $X$ can take on many possible values (think income) or we have few observation for a given value of $X$, we have to write out a more general function.

- These functional forms are unknown which makes life hard.
1. Hypothesis Testing
   - Terminology and Procedure
   - One-Sided Tests
   - Connections
   - Power

2. p-values
   - Mechanics
   - Multiple Testing
   - Fun With Salmon
   - The Significance of Significance

3. What is Regression?
   - Conditional Expectation Functions
   - Nonparametric Regression
   - Best Linear Predictor
   - Ordinary Least Squares

4. 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 $E[Y|X = x]$, i.e. how average income varies with education level.
Nonparametric Regression with Discrete $X$ 

\texttt{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)

Histogram of income
Joint Distribution of $X$ and $Y$ (Income and Education)
Distribution of income given education \( p(y|x) \)
Nonparametric Regression with Discrete $X$
Nonparametric Regression

- This approach 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

- But what do we do when $X$ is continuous and has many values?
- Let’s talk through a few options.
A Binning Approach to CEF for continuous random variables

Estimation of CEF with $n = 10000$ and $n_{\text{cuts}} = 100$
Uniform Kernel Regression: Simple Local Averages

- Dividing into discrete bins can get pretty noisy.
- Another approach is to use a moving local average to estimate $E[Y|X]$.
- We will call this approach **uniform kernel regression** for a reason that will become clear shortly.
Example of Kernel CEF Estimation

Uniform Kernel Estimation

![Graph](image-url)
Example of Kernel CEF Estimation

Uniform Kernel Estimation
Example of Kernel CEF Estimation

Uniform Kernel Estimation

Stewart (Princeton)
Uniform Kernel Regression: Simple Local Averages

- Calculate the average of the observed $y$ points that have $x$ values in the interval $[x_0 - h, x_0 + h]$.
- $h = \text{some positive number (called the bandwidth)}$.
- \textbf{Uniform kernel}: every observation in the interval is equally weighted.

This gives the \textbf{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 $K_h(u) = \frac{1}{2} \mathbf{1}_{\{u \leq 1\}}$.
Changing the Bandwidth

Impact of Bandwidth on Uniform Kernel Estimation
Theorem (Consistency of the Uniform Kernel Density Estimator)

For iid continuous random variables $X_1, X_2, \ldots X_n, \forall x \in \mathbb{R}$,

- if the kernel is uniform, and
- if $h \to 0$ and
- $nh \to \infty$ as
- $n \to \infty$, then

$$\hat{f}_K(x) \xrightarrow{p} f(x).$$

Aronow and Miller Theorem 3.3.8. Proof by weak law of large numbers and the plug-in principle.
The More General Form of the Estimator

**Definition (Kernel Density Estimator)**

Let $X_1, X_2, \ldots X_n$ be iid continuous random variables with common PDF $f$.

Let $K : \mathbb{R} \rightarrow \mathbb{R}$ be a function which is symmetric about the $y$-axis satisfying $\int_{-\infty}^{\infty} K(x)dx = 1$, and let $K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right)$ $\forall x \in \mathbb{R}$ and $h > 0$.

Then a kernel density estimator of $f(x)$ is

$$\hat{f}_K(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i), \forall x \in \mathbb{R}$$

The function $K$ is called the kernel and the scaling parameter $h$ is called the bandwidth.

(Aronow and Miller Definition 3.3.7)
Kernel Estimation

- We can create other estimators by filling in other kernels.
- Generally the idea is to weight things further from the focal point by less than those closer to the focal point.

**Uniform Kernel**

- Calculate the average of the observed $y$ points that have $x$ values in the interval $[x_0 - h, x_0 + h]$.
- Each observation within the interval is given equal weight, each observation outside the interval is given 0 weight.
Kernel Estimation

- We can create other estimators by filling in other kernels.
- Generally the idea is to weight things further from the focal point by less than those closer to the focal point.

**Gaussian Kernel**

- Distance weighted by how far from $x_0$ following the normal density

\[
\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}
\]
Kernel Estimation

- We can create other estimators by filling in other kernels.
- Generally the idea is to weight things further from the focal point by less than those closer to the focal point.
- Triangular
- Distance weighted by how far from $x_0$ using linear distance

![Triangular Kernel Graph]
Kernel Estimation

- We can create other estimators by filling in other kernels.
- Generally the idea is to weight things further from the focal point by less than those closer to the focal point.

- Epanechnikov
- Distance weighted by how far from \( x_0 \) using a parabolic function
Example of Gaussian Kernel

Impact of Bandwidth on Gaussian Kernel Estimation

Stewart (Princeton)
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 choose 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).
1 Hypothesis Testing
   - Terminology and Procedure
   - One-Sided Tests
   - Connections
   - Power

2 p-values
   - Mechanics
   - Multiple Testing
   - Fun With Salmon
   - The Significance of Significance

3 What is Regression?
   - Conditional Expectation Functions
   - Nonparametric Regression
   - Best Linear Predictor
   - Ordinary Least Squares

4 Interpreting Regression
   - Fun With Linearity
Best Linear Predictor

- The CEF can be infinitely complex.
- Estimating the CEF can, therefore, be difficult.
- So, instead we can limit ourselves to classes of functions that approximate the CEF.
  - I.e. we are using a more inflexible estimator.
- In particular, what if we limit ourselves to the class of linear predictors:
  \[ g(X) = a + bX \]

- We call this the linear regression line
- What function minimizes MSE, among this class of functional forms?
Best Linear Predictor

**Theorem**

For a random variable $X$ and $Y$, if $V[X] > 0$, then the best linear predictor (BLP) of $Y$ given $X$ is $g(X) = \alpha + \beta X$ where,

$$\alpha = E[Y] - \frac{Cov[X, Y]}{V[X]}E[X]$$

$$\beta = \frac{Cov(X, Y)}{V(X)}$$

**Corollary**

- The BLP is the best linear predictor of the CEF. I.e. setting $a = \alpha$ and $b = \beta$ minimizes

$$E[(E[Y | X] - (a + bX))^2]$$

- If the CEF is linear, the CEF is the BLP
Best Linear Predictor

Estimation of BLP with $n = 1000$

- BLP
- CEF
- Raw Data

Stewart (Princeton)
Week 4: Testing/Regression
September 21–25, 2020
But what if the CEF isn’t linear?
Well it probably isn’t, but the best linear predictor is still well-defined.
It is the linear projection of $Y_i$ onto $X_i$.
In general, this is distinct from the CEF:
▶ CEF is the best predictor of $Y_i$ among all functions.
▶ Linear projection is the best predictor among linear functions.
The nice thing about the linear projection is that it exists and is well-defined even if the CEF is non-linear.
BLP Approximations Depend on the Marginal Distribution of $X$

Example 2.2.24. ￿e BLP Under Nonlinearity

Suppose that, as in Example 2.2.23, $Y = 10X^2 + W$, $W \sim N(0, 1)$ and $X \sim W$. If $X$ is uniformly distributed between 0 and 1, the CEF of $Y$ given $X$ is the same as above: $E[Y | X] = X^2$. However, we see in the top left panel of Figure 2.2.2 that the BLP approximates the CEF reasonably well only over certain areas. And if the distribution of $X$ changes, the BLP changes. Figure 2.2.2 demonstrates how, with the same (nonlinear) CEF (in black), the BLP (in red) depends on the distribution of $X$.
**Best Linear Predictor**

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

**Figure:** ‘If I fits, I sits’

The BLP is always a line regardless of the data.
1 Hypothesis Testing
   • Terminology and Procedure
   • One-Sided Tests
   • Connections
   • Power

2 p-values
   • Mechanics
   • Multiple Testing
   • Fun With Salmon
   • The Significance of Significance

3 What is Regression?
   • Conditional Expectation Functions
   • Nonparametric Regression
   • Best Linear Predictor
   • Ordinary Least Squares

4 Interpreting Regression
   • Fun With Linearity
Ordinary Least Squares

- Okay, we’ve finally made it to the OLS model you’ve heard so much about!
- We’ve defined a population line of best fit $\beta_0 + \beta_1 X_i$ that approximates the CEF.
- We can now estimate $\beta_0$ and $\beta_1$ like any other population parameters using samples from the joint distribution $f(y, x)(y, x)$.
- The core idea will be to use the plug-in principle. We want the line that minimizes:

$$
(\beta_0, \beta_1) = \arg \min_{\beta_0, \beta_1} E[(Y_i - \beta_0 - \beta_1 X_i)^2]
$$

- So we will plug in the sample analogs to the population:

$$
(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - \bar{\beta}_0 - \bar{\beta}_1 X_i)^2
$$

- This is called the ordinary least squares (OLS) estimator.
Plug-in Estimation of the BLP

Noting we can rewrite \( \frac{\text{Cov}[X,Y]}{V[X]} = \frac{E[XY] - E[X]E[Y]}{E[X^2] - E[X]^2} \), and using the plug-in principle, we can estimate the parameters of the BLP as:

\[
\hat{\alpha} = \bar{Y} - \frac{\bar{X} \cdot \bar{Y}}{\bar{X}^2 - \bar{X}^2} \bar{X} = \bar{Y} - \hat{\beta} \bar{X}
\]

\[
\hat{\beta} = \frac{\bar{X} \cdot \bar{Y}}{\bar{X}^2 - \bar{X}^2} = \frac{\hat{\text{Cov}}(X, Y)}{\hat{V}(X)} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}
\]

This corresponds to the linear projection which minimizes the sum of squared errors.
Fitted linear CEF/regression function

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

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

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

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

Log Settler Mortality vs. Log GDP per capita growth
**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 $X_i$:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

**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

Log GDP per capita growth vs. Log Settler Mortality

\[ r(x) \]

KOR

\[ Y_i \]

\[ \hat{r}(x) \]
Fitted linear CEF/regression function

Log Settler Mortality vs. Log GDP per capita growth

\( r(x) \)

KOR

\( Y_i \)

\( \hat{Y}_i \)

\( \hat{r}(x) \)
Fitted linear CEF/regression function

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

\[ u_i = Y_i - \hat{Y}_i \]

Log GDP per capita growth

Log Settler Mortality

Stewart (Princeton)

Week 4: Testing/Regression

September 21–25, 2020
Why not this line?

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

\[ \hat{u}_i = Y_i - \hat{Y}_i \approx 0 \]
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 mean 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?

![Graph showing comparison of two models, one labeled $\hat{r}_\text{alt}(x)$ and the other $\hat{r}(x)$. The x-axis represents Log Settler Mortality, and the y-axis represents Log GDP per capita growth. The graph compares the residuals of both models across a range of data points.]
Linear Regression: Justification

- The BLP may be a very loose **approximation** to $E[Y|X]$.
- Why would we ever want to do this?
  - Theoretical reason to assume linearity is close enough
  - 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.)
- Perhaps the biggest reason is that it **extends easily** to the case where $X$ is a vector of random variables.
We Covered

- Conditional Expectation Functions
- Nonparametric Regression
- Best Linear Predictors
- Ordinary Least Squares

Next Time: Interpreting Regression
Where We’ve Been and Where We’re Going...

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

- **This Week**
  - hypothesis testing
  - what is regression?
  - nonparametric and linear regression

- **Next Week**
  - inference for simple regression
  - properties of ordinary least squares

- **Long Run**
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference
Hypothesis Testing
- Terminology and Procedure
- One-Sided Tests
- Connections
- Power

p-values
- Mechanics
- Multiple Testing
- Fun With Salmon
- The Significance of Significance

What is Regression?
- Conditional Expectation Functions
- Nonparametric Regression
- Best Linear Predictor
- Ordinary Least Squares

Interpreting Regression
- Fun With Linearity
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 $\beta_0$:

$$E[Y|X = 0] \approx \beta_0 + \beta_1 \cdot 0 = \beta_0$$

2. **Slope**: a one-unit change in $X$ changes our prediction of $Y$ by $\beta_1$

$$E[Y|X = x + 1] - E[Y|X = x] \approx (\beta_0 + \beta_1(x + 1)) - (\beta_0 + \beta_1x)$$

$$= \beta_0 + \beta_1x + \beta_1 - \beta_0 - \beta_1x$$

$$= \beta_1$$
Linear Regression as a Descriptive Model

- A lot of social science reports regression as saying that a one unit change in $X$ is associated with a $\beta$ unit change in $Y$.
- This leans suggestively towards a causal interpretation that if we intervened to change $X$ then $Y$ would change in some way. We will talk more about the assumptions we would need to draw this conclusion later in the semester.
- For now, I find it helpful to think of regression descriptively as talking about different groups of people. Our guess of the mean for units with $x = 2$ is $\beta$ higher than our guess of the mean for the units with $x = 1$.
- This helps clear that it is an approximation to the CEF and that the units being described are different.
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 drawn from the same population we can use information about $X = x_{new}$ to calculate:
    $$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$$
- This prediction is the best in the space of lines in terms of the squared error.
- 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?
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

![Graph showing winning times over years with two models fitting the data.](image-url)
Alternate Models Fit Well, Yield Different Predictions

![Graph showing winning times over years]

- X-axis: Year
- Y-axis: Winning time (s)
- Data points and curves indicate different models yielding different predictions.
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.
- 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

Fivethirtyeight

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/1GsI2u6

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 a Causal Model (A Preview)

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

- For example, can we say that sending that social pressure caused people to vote?

- 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
Fun with Linearity

\[ F(\mu n!) \]

“The Siren’s Song of Linearity”
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

<table>
<thead>
<tr>
<th>Initial data</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Graph 1" /></td>
<td><img src="image2" alt="Graph 2" /></td>
<td><img src="image3" alt="Graph 3" /></td>
<td><img src="image4" alt="Graph 4" /></td>
<td><img src="image5" alt="Graph 5" /></td>
<td><img src="image6" alt="Graph 6" /></td>
<td><img src="image7" alt="Graph 7" /></td>
<td><img src="image8" alt="Graph 8" /></td>
<td><img src="image9" alt="Graph 9" /></td>
<td></td>
</tr>
</tbody>
</table>
We covered

- Some basic insights about how to interpret regression.
- Issues of extrapolation.
- We will return to this more in future weeks.
This Week in Review

- Hypothesis Testing!
- P-Values!
- What is Regression?
- Interpreting Regression!

Going Deeper:


Next week: Properties of Linear Regression with One Explanatory Variable.