MS&E 226: Fundamentals of Data Science

Lecture 13: Additional topics in inference

Ramesh Johari

Warnings

Modeling assumptions: Regression

Remember that most of the inference we have discussed, e.g., for linear regression, requires certain assumptions to hold:

- The population model is linear.
- The errors are normally distributed.
- The errors are i.i.d., with mean zero.

These assumptions are easily violated; as discussed earlier in the course, many of these kinds of violations can be checked by examining the *residuals*, as well as by considering transformations of the variables.

Modeling assumptions: Hypothesis testing

Hypothesis testing has similar warnings attached.

- "Rejecting the null hypothesis" means: the data suggests the null is not true.

Modeling assumptions: Hypothesis testing

Hypothesis testing has similar warnings attached.

- "Rejecting the null hypothesis" means: the data suggests the null is not true.
- We might fail to reject because
  - the null is true; OR
  - because we don't have enough data (low power).

  So failure to reject does not mean the null is true!

Modeling assumptions: Hypothesis testing

Hypothesis testing has similar warnings attached.

- "Rejecting the null hypothesis" means: the data suggests the null is not true.
- We might fail to reject because
  - the null is true; OR
  - because we don't have enough data (low power).

  So failure to reject does not mean the null is true!
- Always be precise about your null hypothesis, and what sampling distribution it implies for your estimator.

  E.g., in linear regression, to get the sampling distribution of $\hat{\beta}_j$ assuming $\beta_j = 0$, we *still* assume the linear normal model with i.i.d. errors.

Association vs. causation

Even in a correct regression model, significant relationships between covariate(s) and the outcome are evidence of *association* (i.e., correlation in the population model), not *causation*.

Suppose we fit a regression with p covariates $X_1, \dots, X_j$, and outcome Y.

A *causal* interpretation of $\hat{\beta}_j$ says that *if* we change X_j by 1 unit, *then* (holding other covariates constant) the outcome Y will change by $\hat{\beta}_j$ units.

Association vs. causation

For us to interpret $\hat{\beta}_j$ causally, we typically want *changes in X_j in our data to be (essentially) independent of other variables* (this is closely related to a concept in statistics called *exogeneity*).
What might go wrong?

Association vs. causation

For us to interpret $\hat{\beta}_j$ causally, we typically want *changes in X_j in our data to be (essentially) independent of other variables* (this is closely related to a concept in statistics called *exogeneity*).
What might go wrong?

- *Collinearity*. If X_j is collinear (correlated with) other covariates in our data, then when X_j varies in our data, the reason for variation in Y might actually be associated variation in the correlated covariates. This can make it difficult to accurately estimate the specific impact of X_j on Y.

Association vs. causation

For us to interpret $\hat{\beta}_j$ causally, we typically want *changes in X_j in our data to be (essentially) independent of other variables* (this is closely related to a concept in statistics called *exogeneity*).
What might go wrong?

- *Collinearity*. If X_j is collinear (correlated with) other covariates in our data, then when X_j varies in our data, the reason for variation in Y might actually be associated variation in the correlated covariates. This can make it difficult to accurately estimate the specific impact of X_j on Y.
- *Omitted variable bias*. If a covariate Z is left out of our model that is correlated with X_j and Y, then when X_j varies in our data, the reason for variation in Y might actually be the associated (unobserved) variation in Z.

Association vs. causation

How do we *enforce* exogeneity? Run an experiment where we (the experimenters) vary X_j exogenously at random, and observe the effect on Y.

This is why *randomized experiments* are the benchmark of causal inference.

More warnings

In the next two sections we look at two additional pitfalls:

- Multiple hypothesis testing
- Post-selection inference

Multiple hypothesis testing

An example: Multiple linear regression

Suppose that I have n rows of data with outcomes $\mathbf{Y}$ and corresponding covariates $\mathbf{X}$. Suppose $p = 100$.

I run a linear regression with all the covariates and check statistical significance. I order the resulting covariates in descending order of p-value:

Covariate index	p-value
40	0.0070
58	0.018
93	0.034
69	0.040
57	0.042
10	0.047

You walk away excited: these six coefficients are all significant at the 95% level, and you now have a starting point for building your model.

An example: Multiple linear regression

In fact: *There is no relationship in this data between* $\mathbf{X}$ *and* $\mathbf{Y}$*!*

I used synthetic data to generate this example, with:

- $Y_i \sim \mathcal{N}(0,1)$ for each i, i.i.d.
- $X_{ij} \sim \mathcal{N}(0,1)$ for each i, j, i.i.d.

So what happened?

What happened?

Recall the p-value is the answer to the following question:

> *What is the chance I would see an estimated coefficient (from the data) as extreme as what I found, if the true coefficient was actually zero?*

What happened?

Recall the p-value is the answer to the following question:

> *What is the chance I would see an estimated coefficient (from the data) as extreme as what I found, if the true coefficient was actually zero?*

If we use a cutoff of 0.05 to determine whether a coefficient is “statistically significant”, then we are willing to accept a 5% rate of *false positives*: coefficients that look large due to random chance, despite the fact that there is really no underlying relationship.

This means with 100 covariates, we should expect 5 of the coefficients to be significant due to random chance alone — even if there is no effect there! (In our case we get slightly more than this.)

Multiple hypothesis testing

This is a systematic issue with statistical significance based on p-values from individual hypothesis tests:

If you use a cutoff of 0.05 (or 0.01, etc.), you should expect 5% (or 1%, etc.) of your discoveries (rejections) to be false positives.

This applies across all hypothesis tests you do: so for example, if you use a 5% cutoff every day at your job on every test you ever run, you will generate false positives in 5% of your hypothesis tests.

Multiple hypothesis testing

Is this a problem? Perhaps not: if you understand that false positives are generated in this way, you can be wary of overinterpreting significance with many hypothesis tests.

The problem is that interpretation of the results becomes much harder: which results are “trustworthy”, and which are “spurious”?

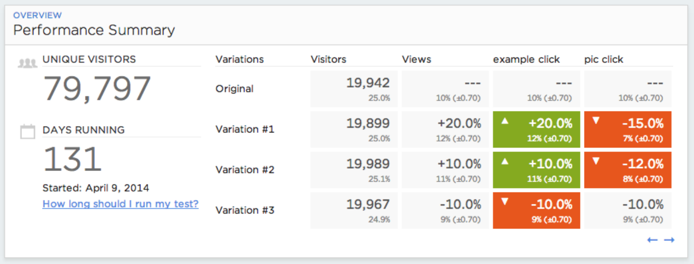
Multiple testing corrections

Multiple testing corrections provide a systematic approach to identifying “meaningful” effects when dealing with many simultaneous hypothesis tests.

This has been an active area of work in the last two decades in statistics, as the range of applications where many hypothesis tests are possible has increased.

Example: The Optimizely results page

Optimizely runs an "A/B testing" platform, a way to test different versions of a website's design against each other. Here is a typical results page for an experiment:



OVERVIEW

Performance Summary

UNIQUE VISITORS

79,797

DAYS RUNNING

131

Started: April 9, 2014

How long should I run my test?

Variations	Visitors	Views	example click	pic click
Original	19,942 25.0%	--- 10% (±0.70)	--- 10% (±0.70)	--- 10% (±0.70)
Variation #1	19,899 25.0%	+20.0% 12% (±0.70)	▲ +20.0% 12% (±0.70)	▼ -15.0% 7% (±0.70)
Variation #2	19,989 25.1%	+10.0% 11% (±0.70)	▲ +10.0% 11% (±0.70)	▼ -12.0% 8% (±0.70)
Variation #3	19,967 24.9%	-10.0% 9% (±0.70)	▼ -10.0% 9% (±0.70)	-10.0% 9% (±0.70)

Real dashboards can have many more simultaneous hypothesis tests present.

Example: Bonferroni correction

The simplest example of a multiple testing correction is the *Bonferroni* correction.

This approach tries to ensure that the probability of declaring even one false positive (also called the *familywise error rate*, FWER) is no more than, e.g., 5%.

Example: Bonferroni correction

The simplest example of a multiple testing correction is the *Bonferroni* correction.

This approach tries to ensure that the probability of declaring even one false positive (also called the *familywise error rate*, FWER) is no more than, e.g., 5%.

The Bonferroni correction declares as significant (rejects the null) any hypothesis where the p-value is $\leq 0.05/p$, where p is the number of hypothesis tests being carried out.

In our example, $p = 100$, so only coefficients with p-values ≤ 0.0005 are declared significant — *none* in the example I showed!

Example: Bonferroni correction

Why does the Bonferroni correction work?

- For a collection of events $A_1, \ldots, A_p$, we have the following bound:

$$\mathbb{P}(A_1 \text{ or } A_2 \text{ or } \cdots A_p) \leq \sum_{j=1}^{p} \mathbb{P}(A_j).$$

Example: Bonferroni correction

Why does the Bonferroni correction work?

- For a collection of events $A_1, \ldots, A_p$, we have the following bound:

$$\mathbb{P}(A_1 \text{ or } A_2 \text{ or } \cdots A_p) \leq \sum_{j=1}^{p} \mathbb{P}(A_j).$$

- So now let A_j be the event that the j'th coefficient is declared significant, with a cutoff of $0.05/p$ on the p-value. Then:

$$\mathbb{P}(A_j | \beta_j = 0) \leq \frac{0.05}{p}.$$

Example: Bonferroni correction

Why does the Bonferroni correction work?

- For a collection of events $A_1, \dots, A_p$, we have the following bound:

$$\mathbb{P}(A_1 \text{ or } A_2 \text{ or } \cdots A_p) \le \sum_{j=1}^{p} \mathbb{P}(A_j).$$

- So now let A_j be the event that the j'th coefficient is declared significant, with a cutoff of $0.05/p$ on the p-value. Then:

$$\mathbb{P}(A_j | \beta_j = 0) \le \frac{0.05}{p}.$$

- Finally, suppose that all the β_j's are in fact zero. Then the probability even one of the A_j's true is $\le p \times 0.05/p = 0.05$.

Example: Benjamini-Hochberg [$*$]

The Bonferroni correction works by essentially forcing your attention only on the smallest p-values (most significant results).

In practice, though, it can be too conservative, especially as the number of hypotheses (e.g., coefficients) increases.

Other methods have emerged to deal with this issue, to allow valid inference while being somewhat less conservative. We consider one, the *Benjamini-Hochberg (BH)* procedure.

Example: Benjamini-Hochberg [$*$]

Suppose we run p hypothesis tests. Of these, suppose that for hypotheses in S_0 the null is in fact true, and on the remainder, the null is false. Suppose that under a given decision procedure, you reject the null for the set of hypotheses in R.

The *false discovery proportion* (FDP) is the fraction of your rejections that were also in the null set:[1]

$$\mathsf{FDP} = \frac{|S_0 \cap R|}{|R|}.$$

The *false discovery rate* (FDR) is the expected value of this fraction over all the randomness in the data:

$$\mathsf{FDR} = \mathbb{E}[\mathsf{FDP}].$$

The BH procedure aims to ensure FDR is less than or equal to α.

[1] FDP is defined to be zero if you make no rejections.

Intuition for false discovery rate [$*$]

Consider the Optimizely results page again.

Suppose we use BH with $\alpha = 0.05$. This ensures that on average, *of those cells that are declared significant*, we will have made mistakes on at most 5% of them.

The criterion is stronger than just controlling each individual test at $\alpha = 0.5$, but weaker than controlling the familywise error rate at $\alpha = 0.05$.

Example: Benjamini-Hochberg [$*$]

The BH procedure at level α is simple to implement:

1. Compute p-values for each of your hypothesis tests, and order them in *increasing* order. Denote these by $q_{(1)}, q_{(2)}, \ldots, q_{(p)}$.
2. Find the *largest* j such that:

$$q_{(j)} \leq \frac{\alpha j}{p}. \tag{1}$$

3. Reject all hypotheses $1, \ldots, j$.

As long as all hypothesis tests are independent of each other, this procedure ensures FDR $\leq \alpha$.[2]

[2] If the hypotheses are not independent, the same result can be guaranteed by changing the right hand side of (1) to $\frac{\alpha j}{p \log p}$.

Example: Benjamini-Hochberg [$*$]

As a numerical example, suppose that we run BH at level $\alpha = 0.05$ with 5 hypothesis tests, where we assume the tests are independent, and we receive p-values $0.001, 0.045, 0.0004, 0.025, 0.15$.

We first order the p-values from lowest to highest: $0.0004, 0.001, 0.025, 0.045, 0.15$.

Since $\alpha/5 = 0.01$, we look for the largest j such that the j'th p-value in the ordered list is $\leq 0.01j$. This is 0.025, so we reject the three hypotheses with p-values $0.0004, 0.001$, and 0.025.

Example: Benjamini-Hochberg [$*$]

The BH procedure is desirable because:

- It is easy to implement.
- It is less conservative than the Bonferroni correction, while still providing useful inference when many hypothesis tests are run.

You should have the habit of always using a procedure like BH when you run many hypothesis tests, to validate that your findings are actually meaningful.

Post-selection inference

Selecting a model

Let's suppose we are building a model using linear regression.

We have $n = 100$ observations $\mathbf{Y}$, two covariates in the design matrix $\mathbf{X}$, and two models we consider:

- In the first model: `Y ~ 0 + X1`.
- In the second model: `Y ~ 0 + X1 + X2`.

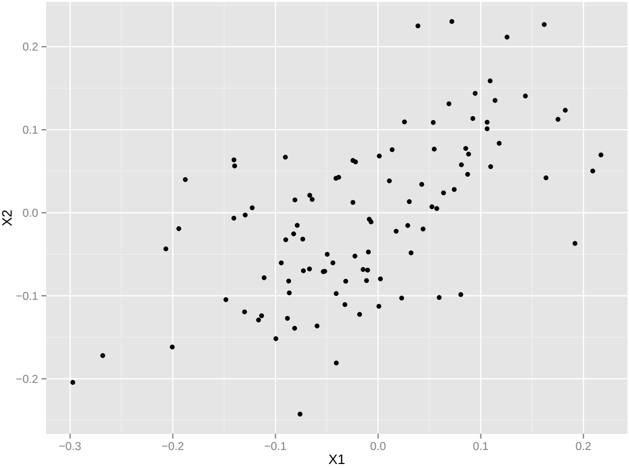
(Ignore the intercept for simplicity.)

Suppose that the design matrix $\mathbf{X}$ is such that the two columns each have norm 1, and their sample correlation is $\approx 2/3$.

Also suppose $Y_i \sim 1.5X_{i2} + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, 1)$ (i.i.d.).

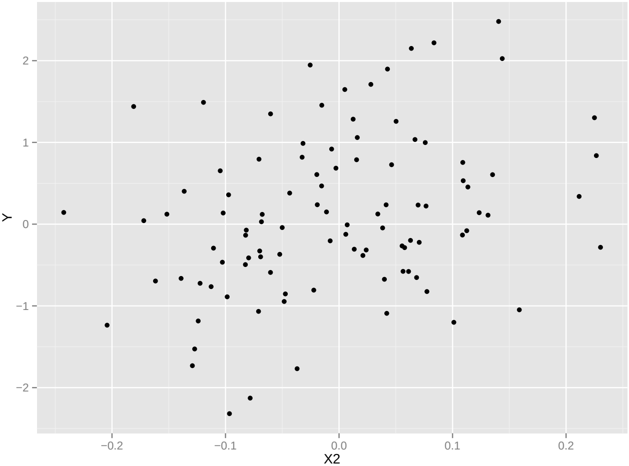
Selecting a model

Example plot of `X2` vs. `X1`:



Selecting a model

Example plot of `Y` vs. `X2`:



Selecting a model

We follow this procedure:

1. We check for significance of $\hat{\beta}_2$ in the complete model. If it is significant, we keep it in the model; if it is not, we drop it.
2. We use the resulting fit to estimate $\hat{\beta}_1$, as well as its standard error $\widehat{\mathsf{SE}}_1$.
3. When done we check if $\hat{\beta}_1$ is statistically significant at the 5% level.

Selecting a model

We do this in 10,000 “parallel universes”, with $n = 100$ in each universe.

In what fraction of our universes *should* $\hat{\beta}_1$ be significant?

Selecting a model

We do this in 10,000 “parallel universes”, with $n = 100$ in each universe.

In what fraction of our universes *should* $\hat{\beta}_1$ be significant?

We find that $\hat{\beta}_1$ is significant in 12.8% of the universes — even though by using a 5% threshold, we should have controlled our false positive rate at 5%!

Selecting a model

What went wrong?

Note that X_1 is highly collinear with X_2. So when we fit the full model, there is a good chance that $\hat{\beta}_2$ might not be significant.

Selecting a model

What went wrong?

Note that X_1 is highly collinear with X_2. So when we fit the full model, there is a good chance that $\hat{\beta}_2$ might not be significant.

But the universes in which this happens are also *more likely* to be those universes where X_1 does a good job of explaining the variation in Y!

Selecting a model

What went wrong?

Note that X_1 is highly collinear with X_2. So when we fit the full model, there is a good chance that $\hat{\beta}_2$ might not be significant.

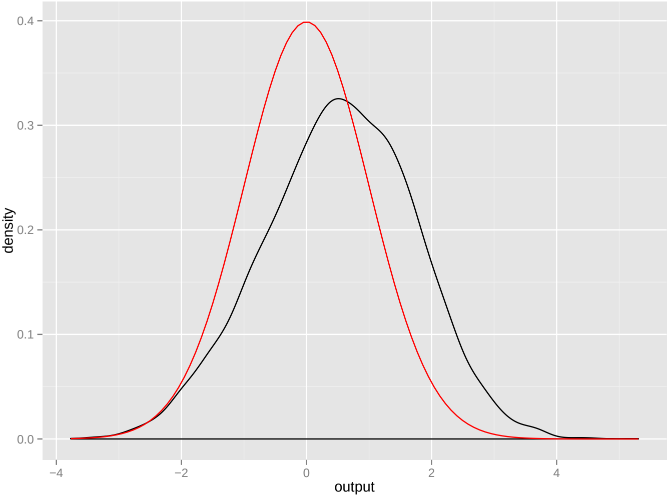
But the universes in which this happens are also *more likely* to be those universes where X_1 does a good job of explaining the variation in Y!

As a result, our selection process makes us more likely to *choose* models in which $\hat{\beta}_1$ is significant, even if the true β_1 is zero; but when we do our hypothesis test, we don't account for this fact.

(Formally, we are doing the hypothesis test assuming that under the null $\hat{\beta}_1$ is $\mathcal{N}(0, \widehat{SE}_1^2)$, when in fact under the null together with our selection process, $\hat{\beta}_1$ is actually more likely to be positive.)

Selecting a model: A picture

The black density plot is the sampling distribution of $\hat{\beta}_1/\widehat{\mathsf{SE}}_1$ under our selection process (over 10,000 parallel universes), while the red curve is the $\mathcal{N}(0,1)$ density:



Post-selection inference

This is an example of *post-selection inference*:

- You apply lots of procedures to your data to find a model you "like".
- *Then* you report your regression table, with p-values, etc.
- But by doing so, you have *favorably biased* your selection of p-values!

(A less generous term for this type of exercise is "p-value hacking".)

Post-selection inference

How can we deal with this?

- We can take into account our selection procedure when we determine the distribution of $\hat{\beta}_1$ under the null hypothesis, and use this "corrected" sampling distribution for significance testing.

  Progress is being made on this approach, but it remains less common in most practical settings.

Post-selection inference

How can we deal with this?

- We can take into account our selection procedure when we determine the distribution of $\hat{\beta}_1$ under the null hypothesis, and use this "corrected" sampling distribution for significance testing.

  Progress is being made on this approach, but it remains less common in most practical settings.
- We can recognize that this type of biasing of our results is an issue, and perhaps take with a grain of salt the low p-values and statistical significance we observe in a final model.

Post-selection inference

How can we deal with this?

- We can take into account our selection procedure when we determine the distribution of $\hat{\beta}_1$ under the null hypothesis, and use this "corrected" sampling distribution for significance testing.

  Progress is being made on this approach, but it remains less common in most practical settings.
- We can recognize that this type of biasing of our results is an issue, and perhaps take with a grain of salt the low p-values and statistical significance we observe in a final model.
- We can *validate* our findings on new data: Fit the same model on a completely new data set, and compute the p-values and significance levels there.

  These are correctly interpretable, because now the model selection is not *based* on the data we are using to check significance.

Concluding thoughts

So now what?

Inference seems loaded with warnings: it is easy to read too much into a statistical analysis, and see “phantom” effects that are not real.

So: how are we meant to build meaningful inferences in practice?

Inference for linear regression

If we've built a meaningful linear regression model, then:

The relationships we declare to be present in our estimated model should be exactly those that are present in the population model.

(*Recall*: Here we are only talking about association, not causation. Additional steps need to be taken to make sure we can interpret a relationship causally.)

External validity

The best evidence that we have a meaningful population model is that it delivers meaningful insight even *outside* of the data on which it was built.

This is the strategy suggested to deal with issues like post-selection inference. Note the similarity to evaluation of predictive models!

External validity

The best evidence that we have a meaningful population model is that it delivers meaningful insight even *outside* of the data on which it was built.

This is the strategy suggested to deal with issues like post-selection inference. Note the similarity to evaluation of predictive models!

What it means in practice is that inference does not stop at just one dataset and one modeling approach; often we should ask for corroborating evidence (from other studies, other datasets, other experiments, common sense, etc.) to strengthen the case for a particular relationship.

Principles to live by

- Be inquisitive.

Principles to live by

- Be inquisitive.
- Be precise.

Principles to live by

- Be inquisitive.
- Be precise.
- Be skeptical.

Principles to live by

- Be inquisitive.
- Be precise.
- Be skeptical.
- Be bold.