Selective inference and regression, part 2

Lucy Gao

February 6, 2024

Recall: agnostic linear regression

$X \in \mathbb{R}^{n \times p+1}$ is a **fixed** covariate matrix with columns $1_n, X_1, \ldots, X_p$, and the response $y \in \mathbb{R}^n$ is a realization from $Y \sim F_Y$, with $Y_1, \ldots, Y_n$ independent and $\mathbb{E}[Y_i] = \mu_i$.

Given a fixed subset of variables $M \subseteq \{1, 2, \ldots, p\}$:

- We focus estimation and inference on $\beta^*_M = (X_M^T X_M)^{-1} X_M^T \mu \in \mathbb{R}^{|M|+1}$, especially the "slope" coefficients
- Scientifically, this means that for each $j \in M$, we want to know about the (approximate linear) association between Y and X_j stratified on values of $\{X_{j'} : j' \neq j\}$

But in practice, we may instead use a variable selection procedure on y to get variables $\hat{M}(y)$. How do we get valid inference on $\beta^*_{\hat{M}(y)}$ using y?

Recall: selective type I error rate

We say that a test controls the selective type I error rate if for any $M \subseteq \{1, 2, \ldots, p\}$, for any $j \in \{1, 2, \ldots, |M|\}$, and for any $0 \leq \alpha \leq 1$, the following holds:

$$F_Y \text{ satisfies } H_0(M, j)) \implies \mathbb{P}_{F_Y}(\text{Reject } H_0(M, j) \text{ using } Y \mid \hat{M}(Y) = M) \leq \alpha$$

"If $H_0(M, j)$ is true, then how often would I choose to reject it, among all repeated experiments (realizations of Y) where the variable selection procedure $\hat{M}(\cdot)$ told me to test $H_0(M, j)$?"

Compare to a context where we commit to testing H_0 without being influenced by the data:

"If H_0 is true, then how often would I choose to reject it, among all repeated experiments (realizations of Y)?"

Unadjusted naive inference

Given y, do inference on $\beta^*_{\hat{M}(y)}$, e.g. test $H_0(\hat{M}(y), j) : [\beta^*_{\hat{M}(y)}]_j = 0$ for $j \in \hat{M}(y)$.

Naively:

- Construct $\hat{\beta}_{\hat{M}(y)} = (X^T_{\hat{M}(y)} X_{\hat{M}(y)})^{-1} X^T_{\hat{M}(y)} y$
- Compare $[\hat{\beta}_{\hat{M}(y)}]_j$ to the large-sample distribution of $[\hat{\beta}_{\hat{M}(y)}(Y)]_j$ under $H_0(\hat{M}(y), j) : [\beta^*_{\hat{M}(y)}]_j = 0$.

(Fudge: actually, divide $[\hat{\beta}_{\hat{M}(y)}]_j$ by the sandwich estimate of the variance of $[\hat{\beta}_{\hat{M}(y)}(Y)]_j$, and compare it to its large sample distribution, which is always $N(0, 1)$ under $H_0(\hat{M}(y), j)$.)

But selective type I error rate control would require us to compare $\hat{\beta}_{\hat{M}(y)}$ to the large-sample *conditional* nulll distribution, $[\hat{\beta}_{\hat{M}(y)}(Y)]_j \mid \hat{M}(Y) = \hat{M}(y)$ under $H_0(\hat{M}(y), j)$.

The naive approach doesn't work

We need to compare $[\hat{\beta}_{\hat{M}(y)}]_j$ to the large-sample *conditional* distribution of

$$[\hat{\beta}_{\hat{M}(y)}(Y)]_j \mid \hat{M}(Y) = \hat{M}(y)$$

under $H_0(\hat{M}(y), j)$; but instead we compared $[\hat{\beta}_{\hat{M}(y)}]_j$ to the large-sample **marginal** distribution of $[\hat{\beta}_{\hat{M}(y)}(Y)]_j$ under $H_0(\hat{M}(y), j)$.

These two distributions cannot be the same unless $\hat{M}(Y)$ is independent of $[\hat{\beta}_{\hat{M}(Y)}(Y)]_j$. (Consider definition of independence and "substitution rule").

Hard to imagine this being true in practice! So in practice, the naive approach will not control the selective type I error rate.

Sample splitting

Take data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ and randomly partition into $\mathcal{D}^{tr} = \{x_i^{tr}, y_i^{tr}\}_{i=1}^{n_{tr}}$ and $\mathcal{D}^{test} = \{x_i^{te}, y_i^{te}\}_{i=1}^{n_{te}}$.

1. Apply a variable selection procedure to $\mathcal{D}^{tr}$ to get $\hat{M}(y^{tr}) \subseteq \{1, 2, \ldots, p\}$.
2. Fit linear regression using the variables in $\hat{M}(y^{tr})$ using $\mathcal{D}^{test}$; report the usual p-values and confidence intervals.

Intuitively, this works because we attack the problem at its root:

- Using the same data for selection and inference creates correlations between the data and the hypothesis that messes up our inference
- *So just use different data for selection and inference*.

Sample splitting: more precisely

Step 1 applies a variable selection procedure to $\mathcal{D}^{tr}$ to get $\hat{M}(y^{tr}) \subseteq \{1, 2, \ldots, p\}$.

This is choosing to focus future inference on:

$$\beta^*_{\hat{M}(y^{tr})}(X^{te}) \equiv [(X^{te}_{\hat{M}(y^{tr})})^T X^{te}_{\hat{M}(y^{tr})}]^{-1} (X^{te}_{\hat{M}(y^{tr})})^T \mu^{te}.$$

Step 2 calculates

$$\hat{\beta}_{\hat{M}(y^{tr})}(X^{te}, y^{te}) \equiv [(X^{te}_{\hat{M}(y^{tr})})^T X^{te}_{\hat{M}(y^{tr})}]^{-1} (X^{te}_{\hat{M}(y^{tr})})^T y^{te},$$

and compares it to the large sample distribution of

$$\hat{\beta}_{\hat{M}(y^{tr})}(X^{te}, Y^{te}) = [(X^{te}_{\hat{M}(y^{tr})})^T X^{te}_{\hat{M}(y^{tr})}]^{-1} (X^{te}_{\hat{M}(y^{tr})})^T Y^{te}.$$

The key to sample splitting

Because Y^{te} is independent of $\hat{M}(Y_{tr})$, the large sample **marginal** distribution of $\hat{\beta}_{\hat{M}(y^{tr})}(X^{te}, Y^{te})$ is the same as its conditional distribution given $\hat{M}(Y_{tr}) = \hat{M}(y_{tr})$.

Because y^{tr} is some fixed realization from Y^{tr}, we know that the large sample marginal distribution of $\hat{\beta}_{\hat{M}(y^{tr})}(X^{te}, Y^{te})$] is our usual Normal distribution. This is what we used for inference in Step 2.

So roughly speaking: whenever we meet the conditions for valid p-values, type I error rate control, and coverage with a fixed hypothesis, we can transfer those guarantees to their selective versions!

Sample splitting works

The implications are:

0. Conditional on selecting any set of variables using the training set, sample splitting produces uniformly distributed p-values whenever F_Y satisfies the selected null hypothesis
1. Sample splitting controls the selective type I error rate
2. Sample splitting attains nominal selective coverage

> "Sample splitting is a simple, radical, almost a-theoretical way to solve the problem of post-selection inference, and as such it appeals to my temperament." - Cosma Shalizi, Notebooks

Illustration: sample splitting works

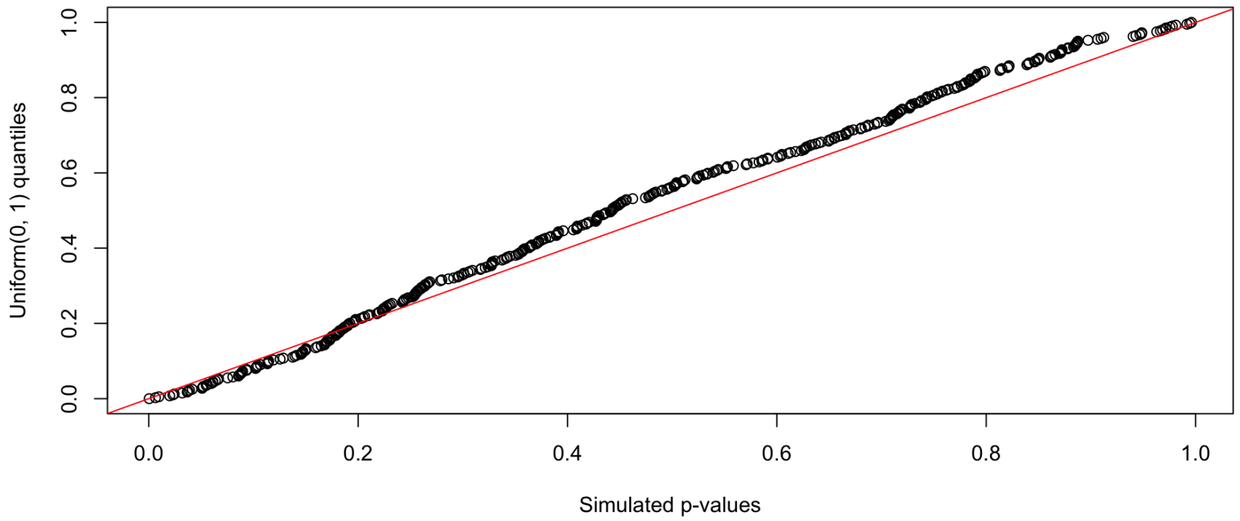
Again generate data with $\mu = \mathbb{E}[Y] = 0_n$, so that regardless of what variables M we pick, our best linear approximation of association should be $(X_M^T X_M)^{-1} X_M^T \mu = 0_{|M|}$, i.e. the selected null hypothesis holds.

```
library(dplyr)

n <- 50
p <- 100
rho <- 0.3

set.seed(1)
Sigma <- (1-rho)*diag(p) + rho*matrix(1, p, p)
X <-  MASS::mvrnorm(n, rep(0, p), Sigma) %>%
    as_tibble(.name_repair = \(x) stringr::str_c("X", 1:p))

do_one_sim <- function(X) {
        df <- X %>% rowwise() %>% mutate(y = rt(1, df=5),
                                         tr_obs = sample(c(T, F), 1))

        df_tr <- df %>% filter(tr_obs)
        df_te <- df %>% filter(!tr_obs)

        # ...
```

Example: sample splitting works

All 400 p-values. Conditionally on each potential set of selected variables, p-value should be approximately Uniform, so we should be plotting 400 Uniform p-values.

```
        qqplot(c(sim_results), qunif(seq(0, 1, length=length(sim_results))),
                    xlab="Simulated p-values", ylab="Uniform(0, 1) quantiles")
        abline(0, 1, col="red")
```



Reflections on sample splitting

Sample splitting really works. As of right now, it's the uncontroversial gold standard for "practical data analysis", e.g. most domain science publications.

Strengths:

- Controls selective type I error rate and attains nominal selective coverage
- Preserves agnostic interpretation of regression
- Can be easily carried out with any variable selection algorithm
- Can be easily extended to generalized linear modelling

Issues:

- Honestly, many! We will go through some …

Issue: target of inference

Sample splitting focuses on:

$$\beta^*_{\hat{M}(y^{tr})}(X^{te}) \equiv [(X^{te}_{\hat{M}(y^{tr})})^T X^{te}_{\hat{M}(y^{tr})}]^{-1}(X^{te}_{\hat{M}(y^{tr})})^T \mu^{te}.$$

Ideally, we'd like to know about:

$$\beta^*_{\hat{M}(y^{tr})}(X) \equiv [(X_{\hat{M}(y^{tr})})^T X_{\hat{M}(y^{tr})}]^{-1}(X_{\hat{M}(y^{tr})})^T \mu.$$

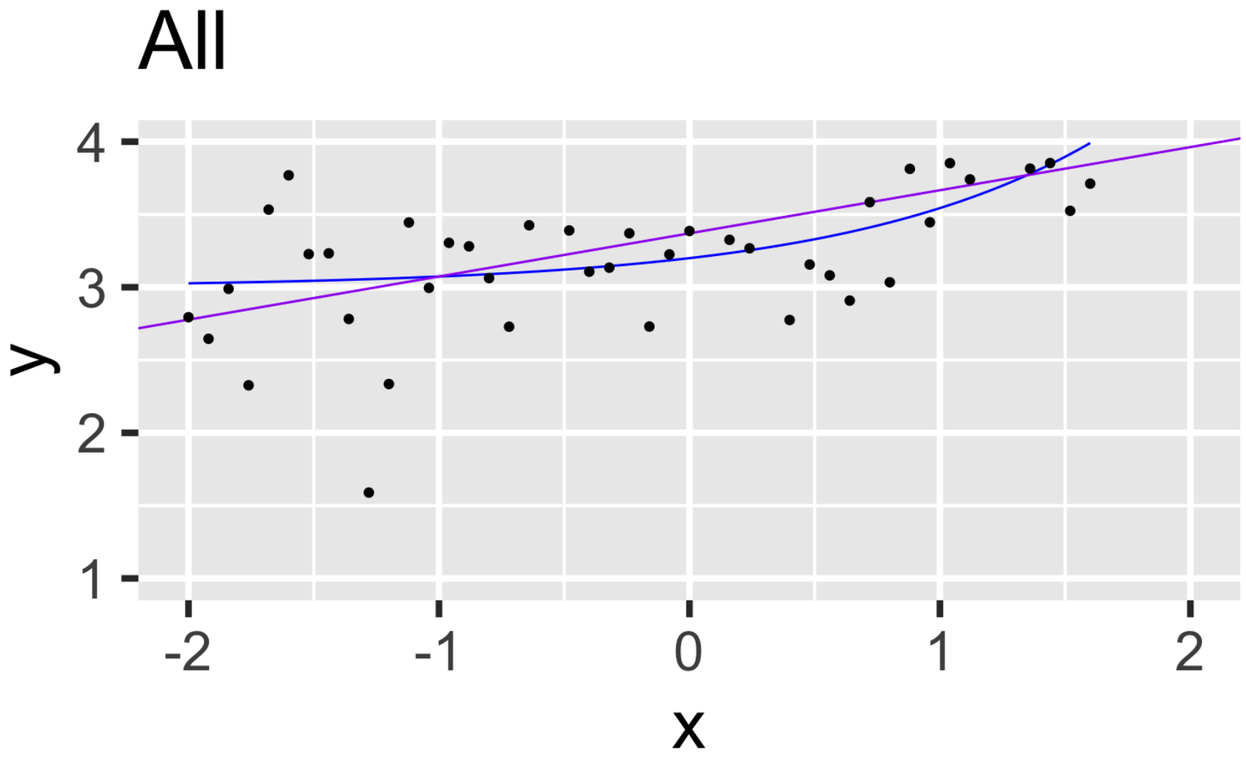
These are not necessarily the same!

Notation looks very similar, but X^{te} has fewer rows than X.

Example: misspecified mean model

$\mathbb{E}[Y_i] = 3 + 0.2 \exp X_i; \beta^*(X) = (3.3705, 0.2965)^T.$

▶ Code

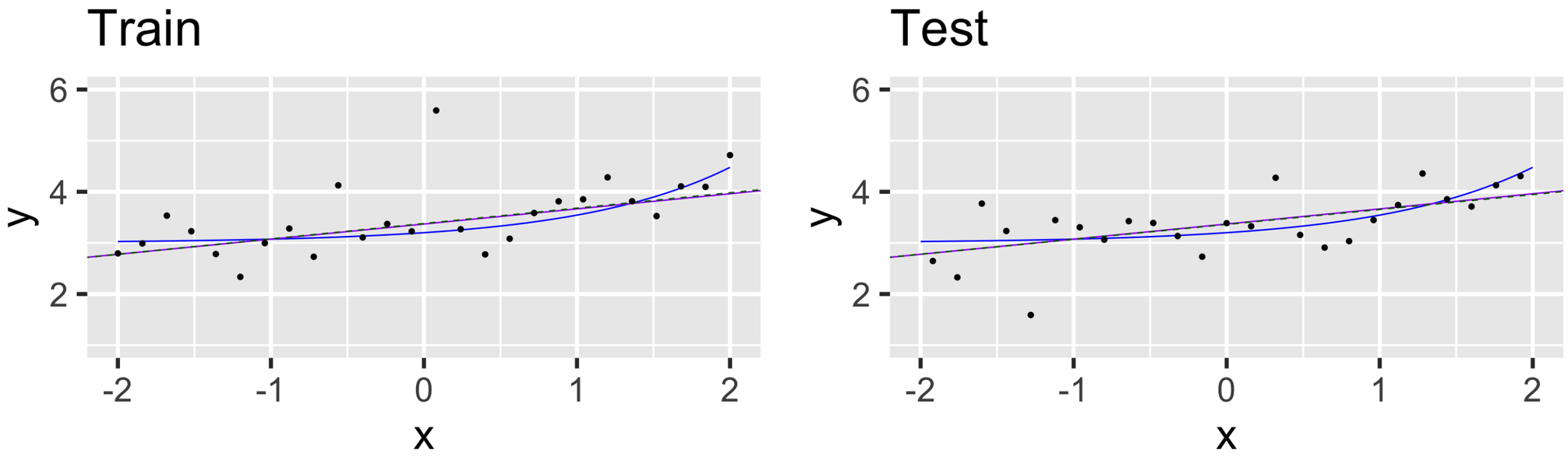


Example: misspecified mean model

If by chance we hit a very balanced split, often little change in target:

$$\beta^*(X) = (3.3705, 0.2965)^T, \beta^*(X^{tr}) = (3.3784, 0.3006), \beta^*(X^{te}) = (3.3623, 0.2919).$$

▶ Code

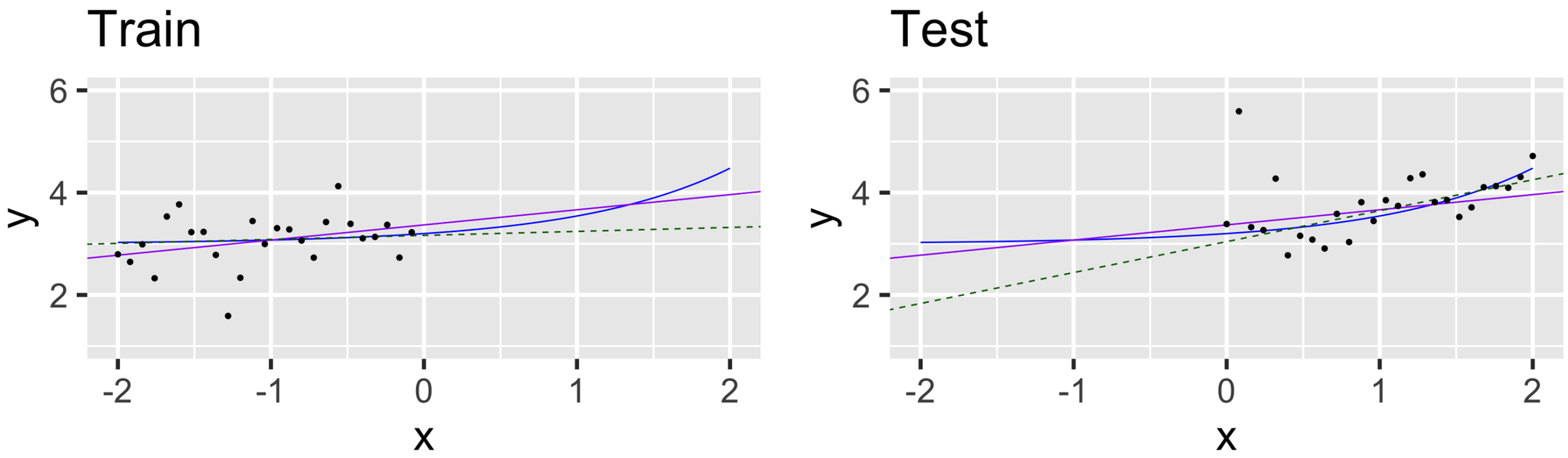


Example: misspecified mean model

If by chance we hit a very imbalanced split, could be a much bigger change in target:

$\beta^*(X) = (3.3705, 0.2965)^T, \beta^*(X^{tr}) = (3.1642, 0.0780). \beta^*(X^{te}) = (3.0424, 0.6046).$

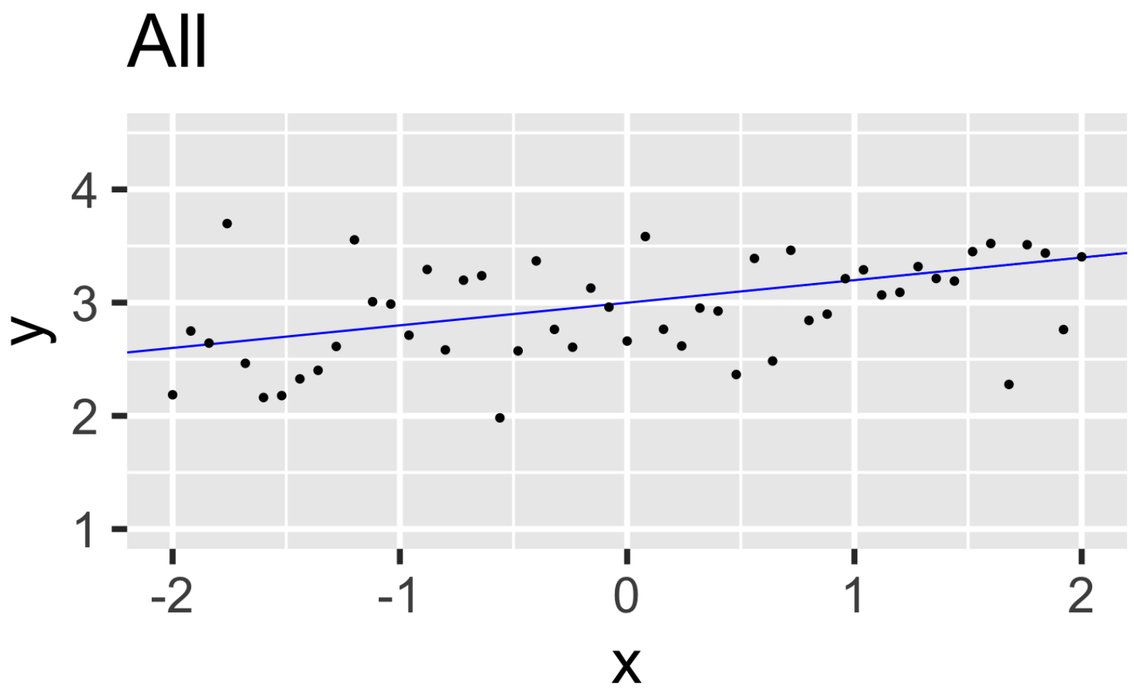
▶ Code



Example: correct mean model

If $\mathbb{E}[Y] = X\beta^*$ and we fit a regression of Y on X, then the target of inference is $(X^TX)^{-1}X^T\beta^* = \beta^*$.

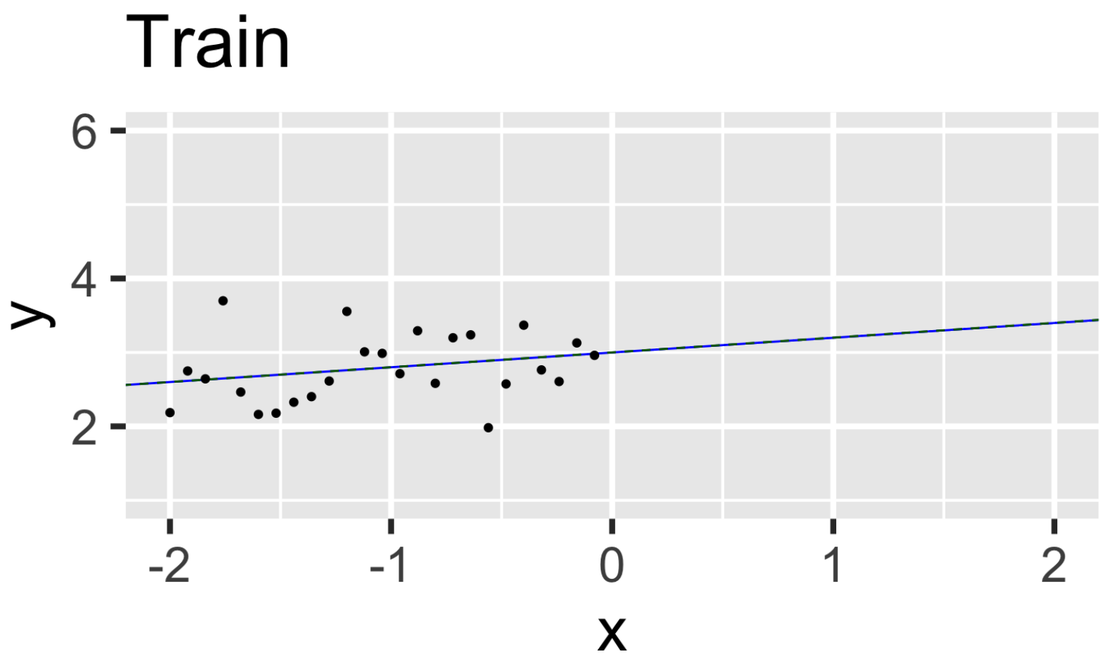
▶ Code

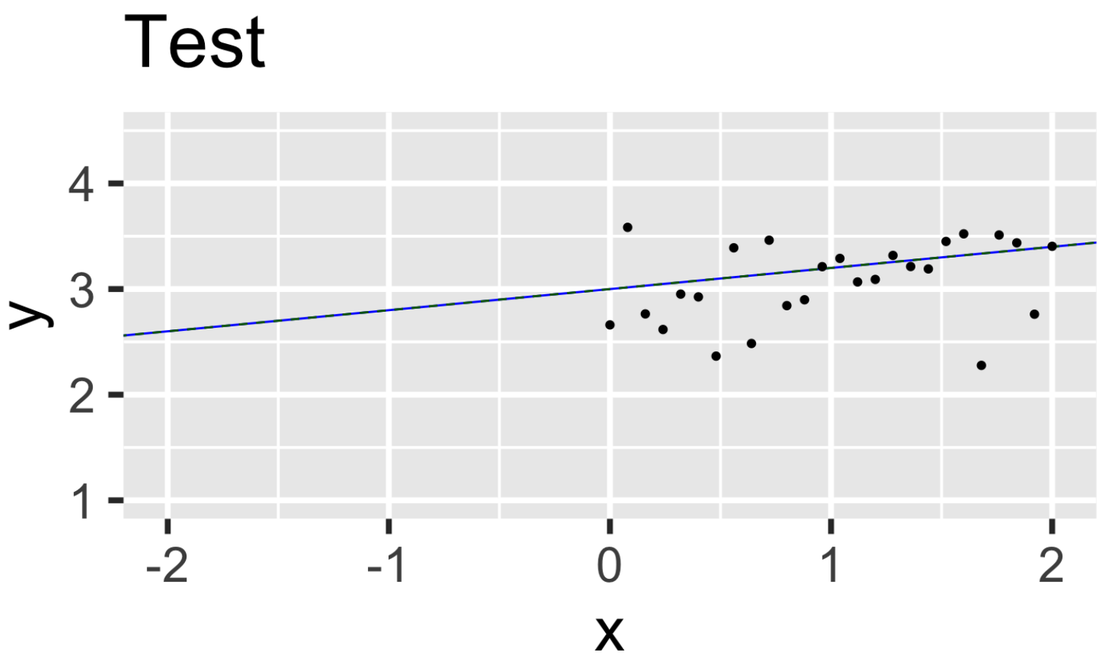


Example: correct mean model

Since $\mathbb{E}[Y] = X\beta^*$, follows that $\mu^{tr} = X^{tr}\beta^*$ and $\mu^{te} = X^{te}\beta^*$. Thus, $\beta^*(X) = \beta^*(X^{tr}) = \beta^(X^{te}) = \beta^*$, **no matter what.**

▶ Code





Implications for variable selection

If we care about finding variables that approximate $\mathbb{E}[Y] = \mu$ well with X (all n observations), then we're definitely ok if X^{tr} looks close to X, even when the true mean model is non-linear.

Otherwise, does it necessarily make sense to use (X^{tr}, y^{tr}) to select variables?

Note that getting balance under uniform random splits should happen often enough, but is far from guaranteed!

Implications for inference

If X^{tr}, X^{te}, and X are all similar looking, then using $\mathcal{D}^{tr}$ to select variables to make inference on in $\mathcal{D}^{te}$ seems reasonable.

But, if we do not select a correct model in the training set, then our inferential target in the test set could be quite different from $\beta^*_{\hat{M}(y^{tr})}(X)$.

- This always happens if the true mean model is non-linear
- This can happen if the true mean model is linear

Again, getting balance under uniform random splits should happen often enough, but is far from guaranteed!

IMO: not a huge issue, but uncomfortable and annoying.

Issue: randomness

We **randomly** split the data into two parts. We already discussed one implication of this; here's another …

Different splits can lead to:

- Different variables selected
- Different p-values, even when same variables are selected

Results are still reproducible since we could set a random seed. But feels like we're leaving something on the table …

“Fix”: cross-validate?

“Obvious” answer is to “cross-validate”:

- Split the data into (say) 5 parts
- For $k = 1, 2, \ldots, 5$:
- Select variables on all but kth part; use kth part for inference

Intuitively, would like to reduce the effect of random variation by combining these p-values. But how???

- The p-values from each of the k folds test different hypotheses
- The p-values from each of the k folds are dependent; combining dependent p-values effectively is difficult, and is an intense subject of current cutting-edge research

… Similar issues, were you to sample split multiple times.

Issue: inefficiency/power loss

By design: for $\epsilon \in (0, 1)$, use $100\epsilon\%$ of the data for variable selection, and use $100(1 - \epsilon)\%$ of the data for inference.

Is your collaborator really going to be happy about having to throw out (say) 50% of the data for inference?

- Remember: more data = higher power
- Remember: if they don't reject the null, they probably can't write a very interesting paper
- "We could not conclusively say that any of these variables are risk factors, but we also can't conclusively say that any of these variables aren't risk factors"

Similarly, is your collaborator really going to be happy about throwing out (say) 50% of the data for variable selection?

Issue: picking split fraction

Ideally: pick the fraction of data to put in the training set so that

1. we have enough data to “do a good job” of variable selection in the training set
2. we have enough data to “do a good job” of inference.

In practice, how do we do this? In fact, in practice, how do we even know that this is possible? (See last slide.)

Often, just based on vibes …

Issue: two sets of estimates

Recall the procedure:

1. Apply a variable selection procedure to $\mathcal{D}^{tr}$ to get $\hat{M}(y^{tr}) \subseteq \{1, 2, \ldots, p\}$.
2. Fit linear regression using the variables in $\hat{M}(y^{tr})$ using $\mathcal{D}^{test}$; report the usual p-values and confidence intervals.

But Step 1 usually produces estimates/predictions - think LASSO or forward stepwise!

Statistically, we can't do anything with them - sample splitting says to throw them out. But they still exist for reviewers and collaborators to ask about …

Can be a real headache for reporting results … especially if you want to do variable selection, prediction, AND inference.

Alternatives to sample splitting

Much statistical research in the last decade has focused on how to characterize and calculate the conditional distribution of

$$\hat{\beta}_{\hat{M}(y)}(Y) \mid \hat{M}(Y) = M.$$

By design, these *conditional selective inference* methods would allow us to construct tests that control the selective type I error rate, and intervals that control the selective coverage.

- See e.g. the methods in the `selectiveInference` R package.

Should you use these methods?

Conditional SI: strengths

Coming from someone who does develop these methods …

Strengths:

- Controls selective type I error rate and attains nominal selective coverage
- We get to use all of the available data for variable selection
- We get to report intervals, p-values, and predictions around a single common set of estimates
- Inference is not based on a random seed
- Methods are available for many commonly used variable selection methods, like LASSO and forward stepwise

Conditional SI: issues

Main issue: **It's really hard to characterize/calculate conditional distribution without further assumptions!**

- Specialized strategies for each variable selection method: "one stats paper at a time" solution
- Need to condition on "more stuff" to eliminate nuisance parameters (see homework problem and next class); even more implications for power
- In fact, current strategies attempt to provide exact, finite-sample inference under a **very strong** assumption: Y_i are independently Normal with **known** covariance
- Can plug in estimate but estimating error variance is notoriously hard (yet more cutting edge research), even if you assume that the linear mean model holds

Also, unclear how much information is used up in variable selection, and so unclear how much information is left to do inference (implications for power)

Practical recommendations

In my opinion:

- If you're doing a data analysis and want to report p-values: either use sample splitting or just don't report them.
- (Awkward middle ground: report them but mention that they aren't valid and should be taken with a grain of salt)
- If you're interested in writing a PhD dissertation, there's a lot of issues you could tackle!
- And in the future, maybe some of the issues with conditional SI will be ironed out

Thursday: a more convincing argument for conditional SI …