Selective inference and regression, part 1

Lucy Gao

January 30, 2024

Randomized controlled trials

Recall: patients randomized to some intervention or control and followed for outcome(s) of interest. Analyze with two-group testing to compare functional of interest.

Strengths

- Gold standard for establishing cause and effect
- Randomization => treatment assignment is not associated with other factors (no confounding)

Issues

- Many exposures/treatments can’t or can’t ethically be studied by random assignment
- Potential lack of generalizability

Observational studies

Many different designs, but common themes:

- No randomization => cannot guarantee no confounding
- Must accept that establishing cause and effect will be challenging, if not outright impossible

Scientific setting:

- Measurements on an outcome of interest (Y) and factors/covariates ($X_1, \ldots, X_p$) of interest
- Primary goal is to infer the association between Y and $X_1, \ldots, X_p$
- Inference (p-values, confidence intervals) **not** optional!

Agnostic linear regression

Let $y = (y_1, \dots, y_n)^T$ be a realization from $Y = (Y_1, \dots, Y_n) \sim F_Y$.

Let $X \in \mathbb{R}^{n \times p}$ be a **fixed** covariate matrix, with $X = [1_n, X_1, \dots, X_p] = [x_1, \dots, x_n]^T$.
(Let $n > p$ for simplicity of notation only.)

We assume that:

- $Y_1, \dots, Y_n$ are mutually independent
- $\mathbb{E}[Y_i] < \infty$ for all i; denote $\mu_i = \mathbb{E}[Y_i], \mu = (\mu_1, \dots, \mu_n)^T$.
- $\mathrm{Var}[Y_i] < \infty$ for all i; denote $\sigma_i^2 = \mathrm{Var}[Y_i]$, and $\Sigma = \mathrm{Cov}(Y) = \mathrm{diag}(\sigma_1^2, \dots, \sigma_n^2)$.

The goal is to study $\mu = \mathbb{E}[Y]$ and its relationship with X.

Linear approximations to μ

We don't know the shape of the relationship between μ and X. But we could always approximate μ with a simple linear function of X:

$$\tilde{\mu} = x_i^T \beta, \quad \tilde{\mu} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_n)^T = X\beta, \qquad \text{for } \beta \in \mathbb{R}^{p+1}$$

The "best" choice of β in terms of squared approximation error is:

$$\beta^* = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} (\mu_i - \tilde{\mu}_i)^2 = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} (\mu_i - x_i^T \beta)^2.$$

Some algebra yields:

$$\beta^* = (X^T X)^{-1} X^T \mu.$$

Picture of best linear approximation

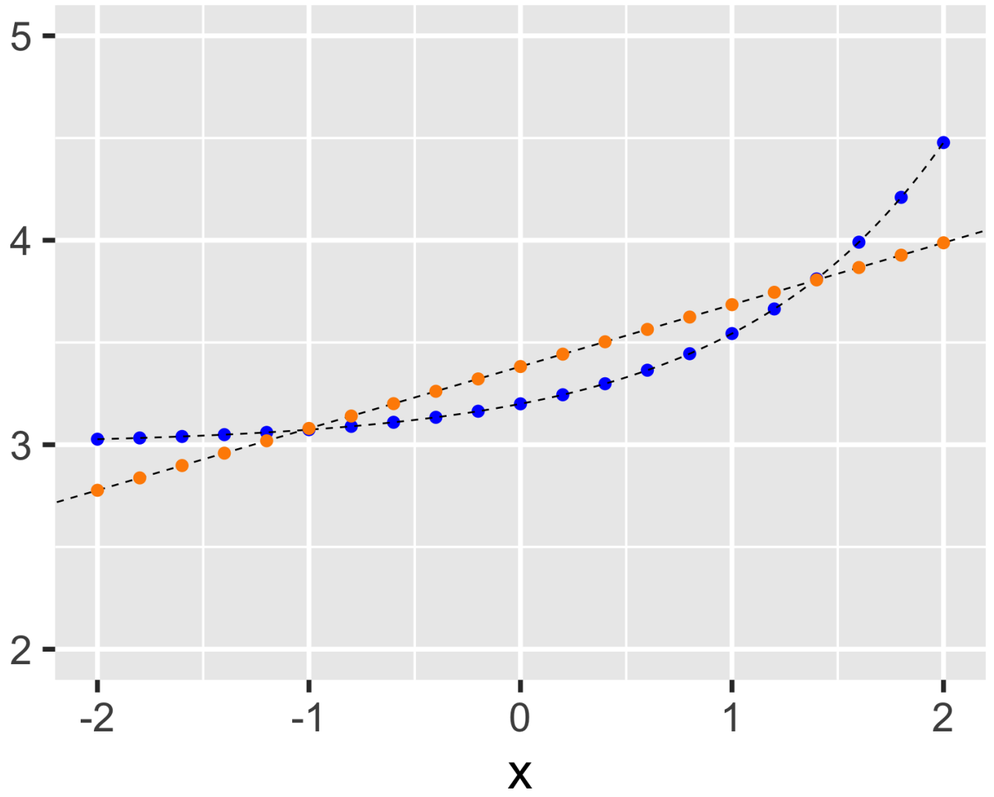
```r
library(ggplot2)
library(dplyr)

df <- tibble(x = seq(-2, 2, length=21)) %>%
    mutate(mu = 3+0.2*exp(x))

beta_star <- broom::tidy(lm(mu~x, data=df)) %>%
    pull(estimate)

df <- df %>% mutate(
    mu_approx = beta_star[1]+beta_star[2]*x
)

ggplot(df) +
    geom_point(aes(x=x, y=mu), colour = "blue",
        size = 3) +
  geom_function(fun = ~3 + 0.2*exp(.x),
    linetype="dashed") +
```



Interpreting intercept

$$x_i^T \beta^* = \beta_0^* + \sum_{j=1}^{p} \beta_j^* [X_j]_i, \quad i = 1, 2, \ldots, n.$$

What is the intercept, β_0^*?

- Imagine a new observation (y', x') with $x' = 0_p$.
- β_0^* is our "best linear approximation" to $\mathbb{E}[Y']$, i.e. the mean outcome of some subpopulation with all covariates $X_1, \ldots, X_p$ set to 0
- Often out of the range of X

Often not scientifically interesting.

Interpreting slopes

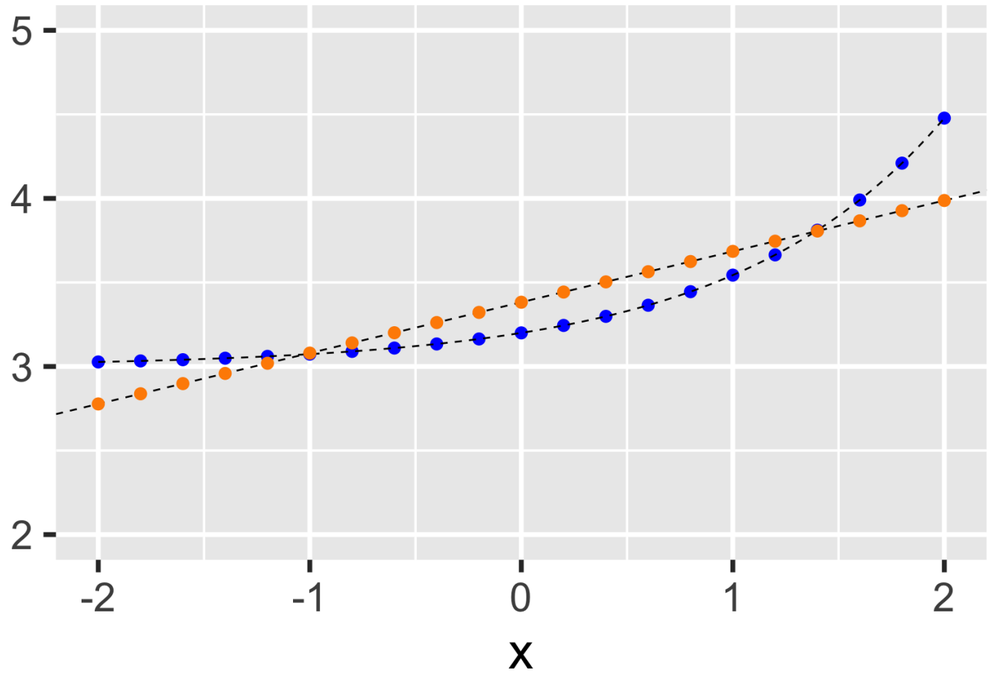
$$x_i^T \beta^* = \beta_0^* + \sum_{j=1}^{p} \beta_j^* [X_j]_i, \quad i = 1, 2, \ldots, n.$$

What is the slope for covariate 1, β_1?

- Imagine two new observations (y', x') and (y'', x''), where x' and x'' differ by one unit in X_1 and are otherwise identical
- β_1 is the "best linear approximation" to $\mathbb{E}[Y''] - \mathbb{E}[Y']$
- $\mathbb{E}[Y''] - \mathbb{E}[Y']$ is the difference in the mean outcome of two subpopulations that differ by one unit in X_1 but agree in their values of $X_2, \ldots, X_p$.

β_1 describes the approximate linear association between Y and X_1 stratified on values of $X_2, \ldots, X_p$.

Back to the picture



- β_1^* here is 0.3; slope of line through orange points
- At $x = -1$, blue point is 3.07
- At $x = 0$, blue point is 3.2; difference = 0.13
- At $x = 1$, blue point is 3.54; difference = 0.34
- At $x = 2$, blue point is 4.48; difference = 0.93

Better approximation in some parts than others, but broadly captures that subpopulations with larger values of X have larger mean outcome.

Estimation of β^*

Recall that:

$$\beta^* = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} (\mu_i - x_i^T \beta)^2 = (X^T X)^{-1} X^T \mu.$$

Consider the minimizer of the plug-in estimate of $\sum_{i=1}^{n} (\mu_i - x_i^T \beta)^2$ that replaces $\mu_i = \mathbb{E}[Y_i]$ with realizations y_i from Y_i:

$$\hat{\beta}(y) = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 = (X^T X)^{-1} X^T y.$$

This is the ordinary least squares estimator (OLSE).

OLSE illustration

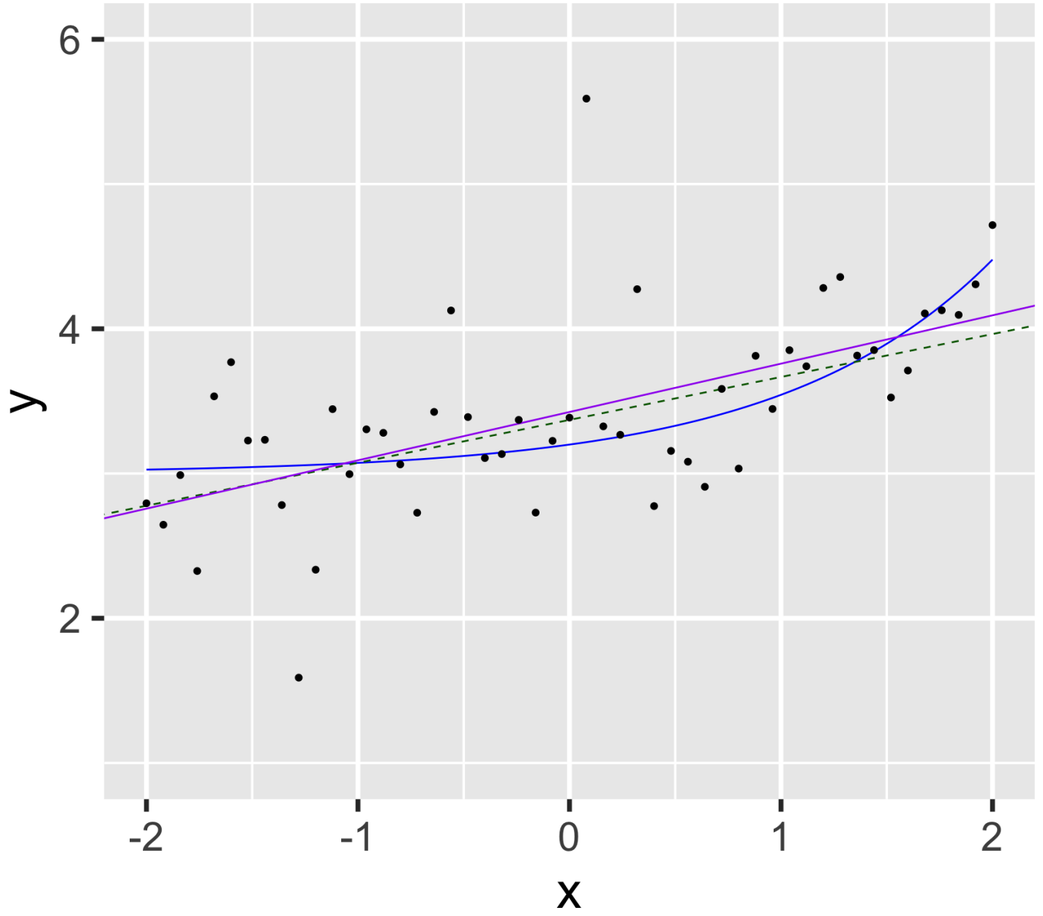
```
library(ggplot2)
library(dplyr)

n <- 51

set.seed(123)
df <- tibble(x = seq(-2, 2, length=n)) %>%
mutate(mu_star = 3+0.2*exp(x),
    y = mu_star + 0.3*rt(n, df=3))

beta_star <- broom::tidy(lm(mu_star~x, data=df))$estima
beta_hat <- broom::tidy(lm(y~x, data=df))$estimate

ggplot(df) +
  xlim(-2, 2) +
    ylim(1, 6) +
      geom_function(fun = ~3 + 0.2*exp(.x), colour="blu
    geom_abline(intercept= beta_star[1], slope=beta_sta
```



Bias and variance of $\hat{\beta}(Y)$

Regardless of F_Y:

$$\mathbb{E}[\hat{\beta}(Y)] = (X^T X)^{-1} X^T \mathbb{E}[Y] = (X^T X)^{-1} X^T \mu = \beta^*$$

The OLSE is unbiased for the best linear approximation (in terms of squared error) to μ.

$$\mathrm{Var}[\hat{\beta}(Y)] = (X^T X)^{-1} X^T \mathrm{Cov}(Y) X (X^T X)^{-1}.$$

This form may be a bit unfamiliar; perhaps helpful to note that when $\mathrm{Cov}(Y) = \sigma^2 I_n$ for some $\sigma^2 > 0$, reduces to

$$(X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}.$$

Inference for β^*

Confidence intervals and p-values are based on the following approximate large-n distribution:

$$\hat{V}(Y)^{-1/2}(\hat{\beta}(Y) - \beta^*) \overset{d}{\approx} N_p(0, I_p),$$

where $\hat{V}(Y)$ is the Huber-White "sandwich" estimator of $\mathrm{Var}(\hat{\beta}_n(Y))$ that replaces diagonal elements of $\mathrm{Cov}(Y) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ with the regression residuals $Y - X\hat{\beta}(Y)$.

Some oversimplifications:

- Haven't specified how mean vector $\mu \in \mathbb{R}^n$ and covariance matrix $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ grows as n increases; regularity conditions omitted
- Technically, under fixed-X framework, sandwich estimator yields asymptotically conservative inference (Fahrmeir 1990)

Inference for β^* in R

- One key change: `sandwich` to get variance-covariance matrix
- Straightforward to implement replacements for test statistic, and p-value calculated with robust standard errors

```r
lm.model <- lm(y~x, data=df)

tidy.lm.summary <- broom::tidy(lm.model) %>% select(term, estimate)
tidy.lm.summary$std.error <- sqrt(diag(sandwich::vcovHC(lm.model)))
tidy.lm.summary %>%
    mutate(statistic = estimate/std.error,
    p.value = 2*pnorm(abs(statistic), lower.tail=FALSE))
```

```
# A tibble: 2 × 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    3.43     0.0773     44.3  0
2 x              0.334    0.0544      6.14 8.45e-10
```

What did we assume?

- We did not have to assume that $\mu(\cdot)$ was linear to say something about F_Y; we make statements about the "best linear approximation" to $\mathbb{E}[Y]$.
- This is a functional of F_Y
- If linearity is badly violated, may have consequences on how much we should care about this functional … eg. consider $\mu(x_i) = x_i^2$
- At no point did we make any assumptions about the parametric family for F_Y (e.g. Gaussian)
- We didn't even make any assumptions about the variance of F_Y
- If linear mean model truly holds (probably doesn't), can drop "best approximation", we are directly learning about $\mathbb{E}[Y]$
- If also Gaussianity and homoskedasticity holds (probably doesn't), then can get more efficient estimates and inference using fully parametric variance estimates

What variables do we use?

Up until now, we have not discussed which covariates we use to approximate $\mu = \mathbb{E}[Y]$. In reality, there are choices to be made!

- We measure p variables $X_1, \dots X_p$
- Within the class of linear approximations, there are 2^p "best linear approximations" we could use, corresponding to subsets of $\{1, 2, \dots, p\}$
- i.e. there are 2^p sets of **functionals** of F_Y we can choose to make inference on …
- and more importantly, use to describe associations between the covariates and the response

What do you do?

The classical, “safe”, paradigm

Pick variables based on **scientific considerations**, and **don’t change your mind after you look at the data**.

Then, the variables we use are not a function of the data realization y and are appropriately viewed as fixed for the inference.

> *“One solution to deciding upon which variables for inclusion in a regression model is to never refine the model for a given dataset. This approach is philosophically pure but pragmatically dubious (unless one is in the context of, say, a randomized experiment) since we may obtain appropriate inference for a model that is a very poor description of the phenomenon under study.”* - Jon Wakefield, in “Bayesian and Frequentist Regression Methods”

Data-driven variable selection

In short: Look at the data then decide on variables to include in your regression.

Examples of formal methods:

- Best subset selection, forward/backwards/stepwise selection
- LASSO, elastic net
- Methods that use causal considerations to pick variables (e.g. estimate a DAG, outcome-adaptive LASSO)

If you include methods that "engineer" features from $X_1, \ldots, X_p$ to use to linearly approximate the mean:

- Tree-based methods (e.g. CART)

Variable selection in science

- 35% did not describe variable selection method in sufficient detail
- 28% used prior knowledge (not data-dependent)
- 37% used a data-dependent variable selection method

Naive approach to inference

If you pick variables with stepwise regression, then R will automatically print p-values for you; and so I promise you that these have been included in some studies.

```
1 summary(best_model_after_forward_stepwise)
```

```
Call:
lm(formula = mpg ~ wt + cyl + hp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9290 -1.5598 -0.5311  1.1850  5.8986 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
wt          -3.16697    0.74058  -4.276 0.000199 ***
cyl         -0.94162    0.55092  -1.709 0.098480 .  
hp          -0.01804    0.01188  -1.519 0.140015    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Naive approach to inference

Two-step procedure:

1. Run a variable selection procedure on the data (X, y) to get a subset of variables $\{1, \ldots, p\}$ to include in our regression; call this $\hat{M}(y)$
2. Infer associations between Y and the variables in $\hat{M}(y)$ by fitting a linear regression of y on the variables in $\hat{M}(y)$ exactly the way we would if the model were fixed and not data-dependent

This is what is printed for stepwise regression; tempting to actively produce such p-values for other methods (e.g. LASSO).

What's wrong with this approach?

The target is data-dependent

Let $M \subseteq \{1, 2, \ldots, p\}$ be fixed.

Let X_M denote the result of subsetting X to the columns in M. Then, modelling Y with the variables in M means the target of estimation and inference is:

$$\beta_M^* = (X_M^T X_M)^{-1} X_M^T \mu$$

Denote $H_0(M) : \beta_M^* = 0$.

The data y is a realization from random variable Y, so $\hat{M}(y)$, the variables selected using y, are a realization from **random variable** $\hat{M}(Y)$.

The hypothesis is data-dependent

This means that given y, you test $H_0(\hat{M}(y))$.

But $H_0(\hat{M}(y))$ is a realization from random variable $H_0(\hat{M}(Y))$!

- If you repeat the study, you may pick a different null hypothesis to test due to random variation in the data collection process
- The variables you chose are not just any old variables - they're specifically ones that seem associated with the particular realization y
- Circular logic: unless we correct for the selection procedure, of course y is going to look associated with variables in $\hat{M}(y)$! (More rigorously, this is because Y and $\hat{M}(Y)$ could be correlated.)

Intuitively, this means that the p-values in Step 2 are generally too small.

Illustration, pt 1

```
n <- 100
p <- 100

library(dplyr)

rho <- 0.3
Sigma <- (1-rho)*diag(p) + rho*matrix(1, p, p)

set.seed(123)
X <-  MASS::mvrnorm(n, rep(0, p), Sigma) %>%
    as_tibble(.name_repair = \(x) stringr::str_c("X", 1

df1 <- X %>% rowwise() %>% mutate(y = 0.5*X1 + 0.2*X3 -

df2 <- X %>% rowwise() %>% mutate(y = 0.5*X1 + 0.2*X3 -

```

```
empty_model1 <- lm(y ~ 1, data = df1)
best_after_fs1 <- step(empty_model1, direction="forwarc
                                                scope = for

empty_model2 <- lm(y ~ 1, data = df2)
best_after_fs2 <- step(empty_model2, direction="forwarc
                                                scope = for
```

```
best_after_fs1$call
```

```
lm(formula = y ~ X1 + X30 + X42, data = df1)
```

```
best_after_fs2$call
```

```
lm(formula = y ~ X1 + X20 + X10, data = df2)
```

Illustration, pt 2

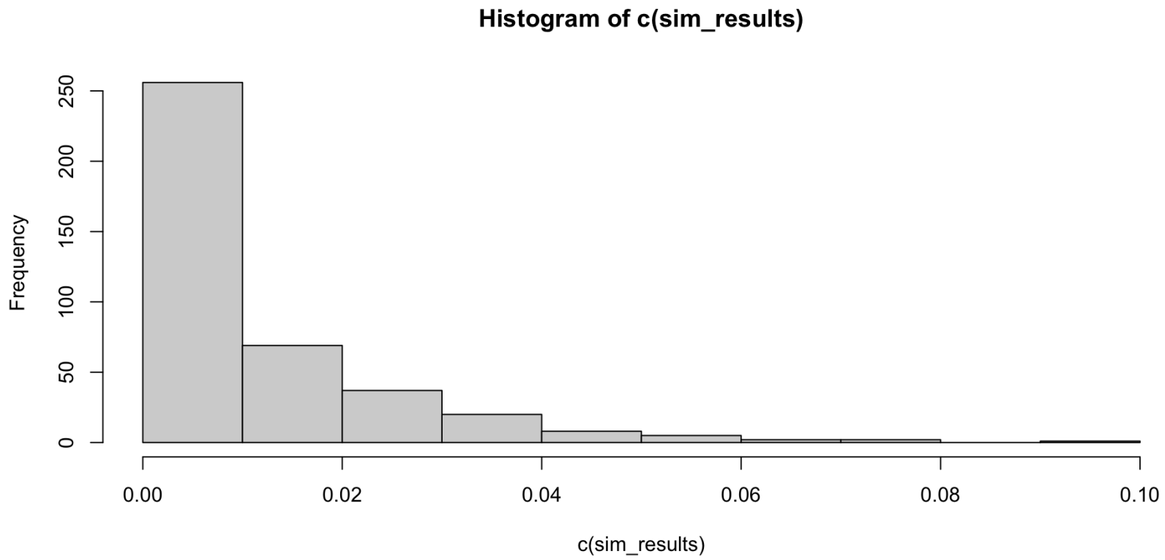
Generate data with $\mu = \mathbb{E}[Y] = 0_n$, so that regardless of what variables M we pick, our best linear approximation of association should be $(X_M^T X_M)^{-1} X_M^T \mu = 0_{|M|}$.

Also, every model we select is "correct"; $\mu = (X_M^T X_M)^{-1} X_M^T \mu = 0_{|M|}$.

```
1  library(dplyr)
2
3  n <- 50
4  p <- 100
5  rho <- 0.3
6
7  set.seed(1)
8  Sigma <- (1-rho)*diag(p) + rho*matrix(1, p, p)
9  X <-  MASS::mvrnorm(n, rep(0, p), Sigma) %>%
10     as_tibble(.name_repair = \(x) stringr::str_c("X", 1:p))
11
12 do_one_sim <- function(X) {
13         df <- X %>% rowwise() %>% mutate(y = rt(1, df=5))
14         empty_model <- lm(y ~ 1, data = df)
15         best_after_fs <- step(empty_model, direction="forward",
16                                            scope = formula(lm(y~., data=df)), steps=2)
17         pvals <- broom::tidy(best_after_fs) %>% pull(p.value)
18
```

Illustration, pt 3

P-values are clearly not what we might hope for them to be.



Also, we certainly aren't happy with how often we reject the null hypothesis!

```
1 mean(sim_results[1, ] <= 0.05 )
```

```
[1] 0.995
```

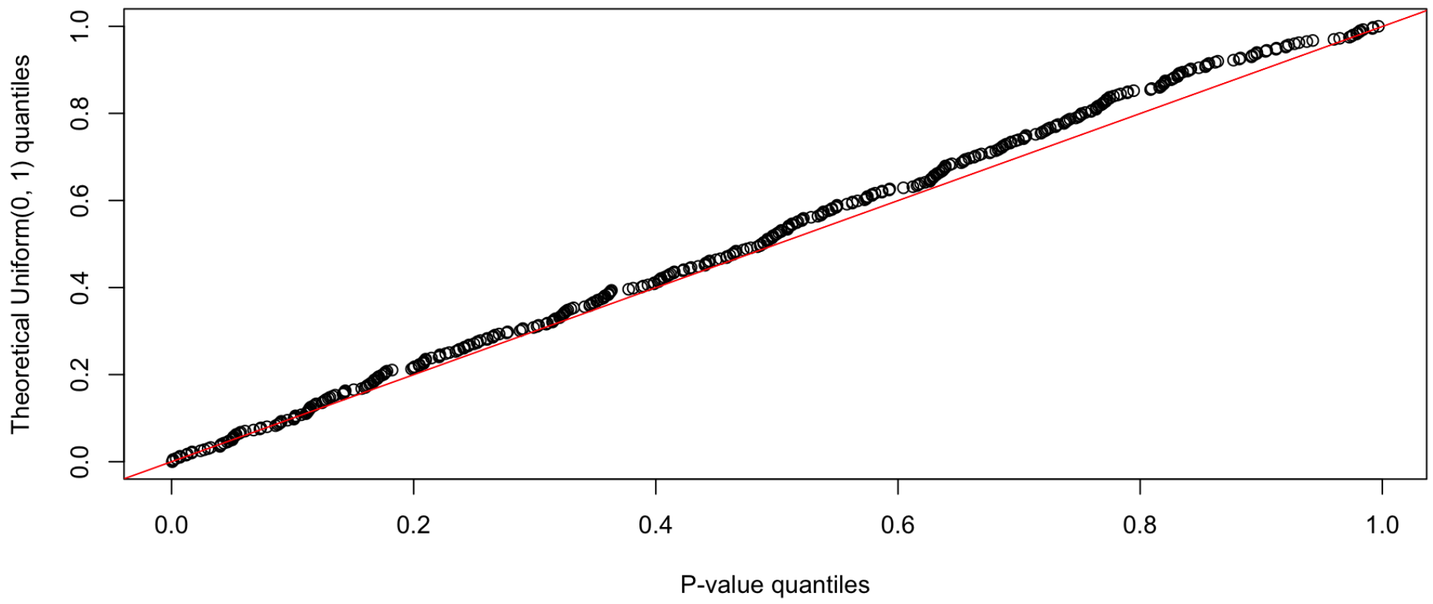
```
1 mean(sim_results[2, ] <= 0.05)
```

```
[1] 0.955
```

Illustration, pt 4

To convince you that this is the effect of selection, here's the exact same thing, except this time I regress Y on X_1 and X_2 on every simulated data set instead of doing model selection:

▶ Code



Selective type I error rate

Let $M \subseteq \{1, 2, \ldots, p\}$ be fixed and arbitrary.

Type I error rate for $H_0(M, j) : (\beta^*_M)_j = 0$, for $j \in |M|$:

$$\mathbb{P}_{F_Y}(\text{Reject } H_0(M, j) \text{ using } Y), \quad \text{where } F_Y \text{ satisfies } [(X_M^T X_M)^{-1} X_M^T \mathbb{E}[Y]]_j = 0$$

Selective type I error rate:

$$\mathbb{P}_{F_Y}(\text{Reject } H_0(M, j) \text{ using } Y \mid \hat{M}(Y) = M), \quad \text{where } F_Y \text{ satisfies } [(X_M^T X_M)^{-1} X_M^T \mathbb{E}[Y]]_j = 0$$

We care about keeping this value below α for any M and j.

Selective type I error rate, in words

Selective type I error rate for $\hat{M}(y) = \{1, 3, 5\}$ and $j = 1$ asks:

- Suppose that we collected data from our study, and based on the data, selected the variables X_1, X_3, and X_5
- Suppose further that there is approximately no linear association between Y and X_1 stratified on values of X_3 and X_5.
- If we **repeat** the study and restrict our attention to only the repetitions (i.e. draws from Y) that also chose to select the variables X_1, X_3, and X_5 …
- How often did we reject the null hypothesis of approximately no linear association between Y and X_1 stratified on values of X_3 and X_5?

Scientific justification

Scientific replication, in my view:

- Consider two studies **that choose the same functional** to address a scientific question
- Do their results agree?

I wouldn't be upset if one study found that the risk ratio was not significantly different from 1 (e.g. Poisson regression), and another study found that the odds ratio (e.g. logistic regression) was significantly different from 1.

When we select a functional from the data, seems natural to only define "mistakes" relative to repeated experiments where we selected the same functional.

> *"The answer must be valid, given that the question was asked."* W. Fithian, D. Sun, and J. Taylor, in "Optimal Inference After Model Selection"

What about confidence intervals?

Recall duality of tests and confidence intervals:

- Can get a confidence interval by calculating the range of nulls you can't reject based on the data
- Can get a hypothesis test by checking if 0 is in the confidence interval

So if the p-values are too small, and the selective type I error rate is not controlled … can expect confidence intervals too short and selective coverage not maintained.

Selective coverage: *[Math Processing Error]*

We care about keeping this value above $1 - \alpha$ for all F_Y.