Just How Easy is it to Cheat a Linear Regression?

Philip Pham

A THESIS

in

Mathematics

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Master of Arts

Spring 2016

Robin Pemantle, Supervisor of Thesis

David Harbater, Graduate Group Chairman

Abstract

As of late, the validity of much academic research has come into question. While many studies have been retracted for outright falsification of data, perhaps more common is inappropriate statistical methodology. In particular, this paper focuses on data dredging in the case of balanced design with two groups and classical linear regression. While it is well-known data dredging has pernicious effects, few have attempted to quantify these effects, and little is known about both the number of covariates needed to induce statistical significance and data dredging's effect on statistical power and effect size. I have explored its effect mathematically and through computer simulation. First, I prove that in the extreme case that the researcher can obtain any desired result by collecting nonsense data if there is no limit on how much data he or she collects. In practice, there are limits, so secondly, by computer simulation, I demonstrate that with a modest amount of effort a researcher can find a small number of covariates to achieve statistical significance both when the treatment and response are independent as well as when they are weakly correlated. Moreover, I show that such practices lead not only to Type I errors but also result in an exaggerated effect size. These findings emphasize the importance of reproducibility and following best practice statistics.

1 Introduction

In a widely cited article, Ioannidis (2005) suggests that most research findings are false due to the prior probability of a true relationship being low, multiple testing by several independent teams, and bias. Ioannidis' claim has generated heated discussion with Jager and Leek (2014) finding that the false discovery rate is merely 14% in top medical literature and refuting that most discoveries are false. However, Gelman and O'Rourke (2014) and Ioannidis (2014) cite problems with bias and methodology in Jager and Leek (2014). Responding to Ioannidis from another angle, Moonesinghe et al. (2007) suggests that replication can solve many of the problems that lead to the false discoveries. Indeed, according to Open Science Collaboration (2015), there is a so-called "replication crisis" in pyschology, for only 36% of replications yielded statistical significance, and the mean effect size was only half as large as the original study. If true, this would be convincing evidence that Ioannidis is correct, but Gilbert et al. (2016) contends that the Open Science Collaboration made errors with procedure, statistical power, and bias. Thus, the debate is still alive and well.

One of the sources of bias that Ioannidis (2005) mentions is selective and distortive reporting. With data dredging, given enough data, one is bound to have statistically significant result. I specifically address how much bias can be obtained from data dredging in the case of balanced design with two groups and classical linear regression. In this paper, data dredging mainly refers to the practice of collecting many variables and running a regression on different subsets of these variables until one obtains the desired result.

While it is well-known data dredging has pernicious effects, few have attempted to quantify these effects, and little is known about both the number of covariates needed to induce statistical significance and data dredging's effect on statistical power and effect size. Permutt (1990) has performed some analysis and simulations on how additional covariates can increase statistical power in the case of covariates that are correlated with response. However, he does not look at the case when the covariates are independent. Senn (1994) looks at how tests of homogeneity can lead to nonrandom assignment of the treatment. If the correlations between the covariates and response are known beforehand, manipulating assignment of the treatment is equivalent in spirit to mining covariates, and he gives an algorithm for the researcher to pick the treatment group to increase the probability of statistical significance while maintaining "balanced" treatment and control groups. In practice, these correlations are usually not known. He does actually mention the case of collecting additional uncorrelated covariates, but only at a philosophical level, and terms it “post-study anxiety.” To avoid problems with multiple hypothesis testing and bias, he suggests true random assignment, specifying the model before the experiment, and only looking at additional covariates post-study to inform future models. Despite the recommendations, he does not provide anything quantitative about how easily data dredging can be used to cheat and favor the treatment.

I look at the effects of data dredging from a few different angles:

1. First, I look at the problem from a theoretical perspective, where we have an unbounded supply of covariates. Here, I show that we can induce statistical significance and make the effect size as large as possible.

2. Next, I look at the case where the treatment and the response are independent by simulation. I find on average how many independent covariates are needed for statistical signficance.

3. Thirdly, I look at the case when the effect size is small, so the statistical power is low, and therefore, the effect is hard to detect. I analyze how data dredging can give a false sense of reproducibility and exaggerate the magnitude of the effect.

2 Unbounded Supply of Covariates

Consider an experiment with $2N$ subjects, where we assign N subjects to a treatment group, and N subjects to the control group. We observe the response $\mathbf{Y}$ along with covariates $\{\mathbf{z}_j : j = 1, 2, \ldots\}$. Let $\mathbf{X}$ represent whether the subject was assigned to a treatment group or not. These vectors will be of length $2N$ with each entry corresponding to a subject.

To see whether the treatment was effective or not, we model the relationship between $\mathbf{Y}$, $\mathbf{X}$, and some subset of $S \subset \{\mathbf{z}_j : j = 1, 2, \ldots\}$ to be linear, where we assume that there is normally distributed noise in the observation. Let $S = \{\mathbf{z}_{j_1}, \ldots, \mathbf{z}_{j_{m-1}}\}$, so

$$\mathbf{Y} = \beta^S_{\mathbf{X}}\mathbf{X} + \sum_{k=1}^{m-1} \beta^S_k \mathbf{z}_{j_k} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$.

Let $\boldsymbol{\beta}_S = (\beta^S_1, \ldots, \beta^S_{m-1}, \beta^S_{\mathbf{X}})$. The maximum likelihood estimate for $\boldsymbol{\beta}_S$ is

$$\hat{\boldsymbol{\beta}}_S = \begin{pmatrix} \hat{\beta}^S_1 \\ \vdots \\ \hat{\beta}^S_m \\ \hat{\beta}^S_{\mathbf{X}} \end{pmatrix} = (\mathbf{Z}_S^\intercal \mathbf{Z}_S)^{-1} \mathbf{Z}_S^\intercal \mathbf{Y}, \tag{2.2}$$

where $\mathbf{Z}_S$ is a $2N \times m$ matrix with the vectors of S as the first $m-1$ columns and $\mathbf{X}$ as the last column (Bickel and Doksum, 2015).

If we restrict $\mathbf{X}$ and the $\mathbf{z}_j$s to be vectors of 0s and 1s a large $\hat{\beta}^S_{\mathbf{X}}$ suggests that the effect size is large. On the other hand if $\mathbf{X}$ and $\mathbf{Y}$ are independent, we would expect that $\hat{\beta}^S_{\mathbf{X}}$ is close to 0.

I show that if one collects data forever and has an infinite set of independent $\{\mathbf{z}_j\}$, one can make the estimate $\hat{\beta}^S_{\mathbf{X}}$ as large as possible even if the actual value is $\beta^S_{\mathbf{X}} = 0$, that is, $\mathbf{X}$ and $\mathbf{Y}$ are independent.

Let there be $2N$ subjects. We randomly assign half the subjects to a treatment group. Let $\mathbf{X} = (X_1, \ldots, X_{2N}) \in \mathbb{R}^{2N}$ be defined

$$X_i = \begin{cases} 1 & \text{subject } i \text{ is assigned to the treatment group} \\ 0 & \text{subject } i \text{ is in the control group.} \end{cases} \tag{2.3}$$

For each subject, we observe an independent response $Y_i \sim \mathcal{N}(0, 1)$ that is independent of X_i. Define $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_{2N}) \sim \mathcal{N}(\mathbf{0}, I)$. Suppose we have a infinite sequence of independent random vectors $\{\mathbf{z}_j\}$, where $\mathbf{z}_j = (z_{1,j}, \ldots, z_{2N,j})^\intercal \in \mathbb{R}^{2N}$ and $z_{i,j} \sim$ Bernoulli(1/2).

Theorem 2.1. *Let $S = \{\mathbf{z}_{j_1}, \ldots, \mathbf{z}_{j_{m-1}}\}$ be a subset of $\{\mathbf{z}_j\}$ such that $S \cup \{\mathbf{X}\}$ is linearly independent. Define $\mathbf{Z}_S$ to be a $2N \times m$ matrix, where the vectors of S make up the first $m-1$ columns, and $\mathbf{X}$ is the last column, where m can be any nonnegative integer. Define $\hat{\boldsymbol{\beta}}_S$ as in Equation 2.2.*

Then, for any $M > 0$ and $\epsilon > 0$, there exists some n such that $N \geq n$ implies that there exists S such that $\mathbb{P}\left(\hat{\beta}^S_{\mathbf{X}} > M\right) > 1 - \epsilon$. Moreover, if $s^2 = \dfrac{|\mathbf{Y} - \mathbf{Z}_S\hat{\boldsymbol{\beta}}_S|^2}{2N - m}$, for any $\delta > 0$ and level of significance $\alpha > 0$,

$$\mathbb{P}\left(\frac{\hat{\beta}^S_{\mathbf{X}}}{s\sqrt{\left((\mathbf{Z}_S^\intercal \mathbf{Z}_S)^{-1}\right)_{mm}}} \geq T^{-1}_{2N-m}(1 - \alpha/2)\right) > 1 - \delta,$$

where T_{2N-m} is the cumulative distribution function for the t distribution with $2N - m$ degrees of freedom. That is, our estimate for $\hat{\beta}^S_{\mathbf{X}}$ will be statistically significant according to a two-sided t-test.

First, let us establish a few linear algebra facts.

Lemma 2.2. *Let $\mathbf{Z}$ be an $n \times m$ matrix of full rank, where $m < n$. $\mathbf{Z}^\intercal\mathbf{Z}$ is an invertible $m \times m$ matrix. Now, define $\mathbf{Z}_{(k)}$ to be the $nk \times m$ matrix, where the rth row of $\mathbf{Z}_{(k)}$ is the $\lceil r/k \rceil$th row of $\mathbf{Z}$. $\mathbf{Z}_{(k)}^\intercal \mathbf{Z}_{(k)}$ is also invertible and its inverse is $k^{-1}\left(\mathbf{Z}^\intercal\mathbf{Z}\right)^{-1}$.*

Proof. If $\mathbf{Z}$ has rank m, then $\mathbf{Z}$ has trivial null space by rank-nullity theorem. Let $\mathbf{x} \in \mathbf{R}^m$. $\mathbf{Z}$, so

$$\mathbf{Z}^\intercal\mathbf{Z}\mathbf{x} = \mathbf{0} \Leftrightarrow \mathbf{0} = \mathbf{x}^\intercal\mathbf{Z}^\intercal\mathbf{Z}\mathbf{x} = (\mathbf{Z}\mathbf{x})^\intercal(\mathbf{Z}\mathbf{x}) \Leftrightarrow \mathbf{Z}\mathbf{x} = \mathbf{0}$$

since the dot product is a norm. Thus, $\mathbf{Z}^\intercal\mathbf{Z}$ is an $m \times m$ matrix with trivial null space, so it is invertible.

Now, clearly $\mathbf{Z}_{(k)}$ will have full rank, too, so $\mathbf{Z}_{(k)}^{\mathsf{T}}\mathbf{Z}_{(k)}$ is an invertible $m \times m$ matrix, too.

$$\begin{aligned}\left(\mathbf{Z}_{(k)}^{\mathsf{T}}\mathbf{Z}_{(k)}\right)_{ij} &= \sum_{l=1}^{nk}\left(\mathbf{Z}_{(k)}^{\mathsf{T}}\right)_{il}\left(\mathbf{Z}_{(k)}\right)_{lj} \\ &= \sum_{p=0}^{n-1}\sum_{l=1}^{k}\left(\mathbf{Z}_{(k)}^{\mathsf{T}}\right)_{i,pk+l}\left(\mathbf{Z}_{(k)}\right)_{pk+l,j} \\ &= \sum_{p=0}^{n-1}\sum_{l=1}^{k}\left(\mathbf{Z}^{\mathsf{T}}\right)_{i,p+1}\left(\mathbf{Z}\right)_{p+1,j} = k\sum_{p=1}^{n}\left(\mathbf{Z}^{\mathsf{T}}\right)_{ip}\left(\mathbf{Z}\right)_{pj} \\ &= k(\mathbf{Z}^{\mathsf{T}}\mathbf{Z})_{ij},\end{aligned}$$

so $\mathbf{Z}_{(k)}^{\mathsf{T}}\mathbf{Z}_{(k)} = k\mathbf{Z}^{\mathsf{T}}\mathbf{Z}$, which implies that $\left(\mathbf{Z}_{(k)}^{\mathsf{T}}\mathbf{Z}_{(k)}\right)^{-1} = k^{-1}\left(\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\right)^{-1}$. □

Lemma 2.3. *Consider the matrices* $\mathbf{F}_n$ *defined as follows. Let* $\mathbf{F}_1 = \begin{pmatrix}1\end{pmatrix}$. *For* $n > 1$, *define the* $n \times n$ *matrix* $\mathbf{F}_n$ *recursively by*

$$(\mathbf{F}_{n+1})_{i,j} = \begin{cases} (\mathbf{F}_n)_{i-1,j-1} & i > 1,\ j > 1 \\ 1 & i = 1,\ j \in \{1, 2, N+1\} \\ 1 & j = 1,\ i \equiv 1 \pmod 2 \\ 0 & \textit{otherwise.} \end{cases} \tag{2.4}$$

The last row of the inverse of this matrix is

$$\begin{pmatrix}-F(1) & -F(2) & \cdots & -F(n-2) & F(n-2) & F(n-1)\end{pmatrix},$$

where $F(k)$ *is the* k*th Fibonnaci number, where* $F(0) = 1$, $F(1) = 1$, *and* $F(k) = F(k-1) + F(k-2)$ *for* $k > 1$.

Proof. $\mathbf{F}_n^{-1}\mathbf{F}_n$ is the identity. Define

$$\mathbf{r} = \begin{pmatrix}-F(1) & -F(2) & \cdots & -F(n-2) & F(n-2) & F(n-1)\end{pmatrix}.$$

Let $C_1, \ldots, C_n$ be the columns of $\mathbf{F}_n$. We show that $\mathbf{r}C_j = 0$ for all $1 \leq j < n$, and $\mathbf{r}C_n = 1$.

First by induction, we have that for $n \geq 2$

$$F(k) = F(k-1) + F(k-2) = \left(F(0) + \sum_{j=1}^{n-3} F(j) \right) + F(k-2) = 1 + \sum_{j=1}^{n-2} F(j). \tag{2.5}$$

Now, using this, for the last column:

$$\mathbf{r}C_n = -\sum_{j=1}^{n-3} F(j) - F(n-2) + F(n-2) + F(n-1) = 1.$$

Looking closely at the recursive relationship in Equation 2.4, we have that for $j < n$,

$$(\mathbf{F}_n)_{i,j} = \begin{cases} 1 & j = i+1 \\ 1 & j < i \text{ and } j \equiv i \pmod 2 \\ 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

Therefore, for $j < n$, we have that

$$\begin{aligned} \mathbf{r}C_j &= \begin{cases} -F(j-1) - \displaystyle\sum_{k=0}^{(n-3-j)/2} F(j+2k) + F(n-2), & j \text{ is odd} \\ -F(j-1) - \displaystyle\sum_{k=0}^{(n-2-j)/2} F(j+2k) + F(n-1), & j \text{ is even.} \end{cases} \\ &= \begin{cases} -F(j-1) - \displaystyle\sum_{k=0}^{(n-5-j)/2} F(j+2k) + F(n-4), & j \text{ is odd} \\ -F(j-1) - \displaystyle\sum_{k=0}^{(n-4-j)/2} F(j+2k) + F(n-3), & j \text{ is even.} \end{cases} \end{aligned}$$

since $F(k) - F(k-1) = F(k-2)$. For the odd case, We can do this reduction a total of $(n-3-j)/2+1$ times, so we find that

$$F(n-2) - \sum_{k=0}^{(n-3-j)/2} F(j+2k) = F\left(n - 2 - 2\left(\frac{n-3-j}{2} + 1\right)\right) = F(j-1).$$

For the even case,

$$F(n-1) - \sum_{k=0}^{(n-2-j)/2} F(j+2k) = F\left(n - 1 - 2\left(\frac{n-2-j}{2} + 1\right)\right) = F(j-1).$$

Thus, $\mathbf{r}C_j = 0$ for $j < n$. □

With Lemmas 2.2 and 2.3, we can now prove Theorem 2.1.

Proof. Without loss of generality reindex our vectors, so that

$$X_1 = \cdots = X_N = 1 \text{ and } X_{N+1} = \cdots = X_{2N} = 0.$$

Also, index it so that

$$Y_1 \leq Y_2 \leq \cdots \leq Y_N.$$

Now, $\mathbb{P}(\mathbf{z}_j = (a_1, \ldots, a_{2N})) = 2^{-2N}$, so by independence and Borel-Cantelli lemma, every possible vector of 0s and 1s occurs infinitely often (Durrett, 2010).

Now, suppose $N = r2^p$, and let $m = r2^{p-q}$, where $r, p, q \in \mathbf{N}$, and $q \leq p$. Define $\mathbf{e}_j^*$ to be vectors such that

$$\left(\mathbf{e}_j^*\right)_i = \begin{cases} 1, & \lceil i/2^q \rceil = j \\ \\ 0, & \text{otherwise.} \end{cases}$$

That is, we segment our data into groups of size 2^q.

We can pick S to be a subset of $\{\mathbf{z}_j\}$ such that $\mathbf{Z}_S$ is an arbitrary matrix of 0s and 1s in the first m columns such that the columns span $\operatorname{span}(\mathbf{e}_1^*, \ldots, \mathbf{e}_m^*)$. Note that the last column of $\mathbf{Z}_S$ is fixed to have 1s in the first N rows and the rest 0. All the rows after the Nth row of $\mathbf{Z}_S$ are 0.

Now, recall that $\hat{\boldsymbol{\beta}}_S$ minimizes

$$(\mathbf{Y} - \mathbf{Z}_S\hat{\boldsymbol{\beta}}_S)^\intercal(\mathbf{Y} - \mathbf{Z}_S\hat{\boldsymbol{\beta}}_S),$$

and the mean is the best estimator in the case that the only covariate is a vector of all 1s, that is, we are only estimating the intercept.

Define $\overline{\mathbf{Y}}_{i,j}$ to be $\sum_{k=i}^{j} Y_i/(j-1+1)$. Having covariates $\{\mathbf{e}_1^*, \ldots, \mathbf{e}_m^*\}$ is equivalent to estimating the intercept in groups of size 2^q.

Let $\mathbf{Z}$ be the $2N \times m$ matrix with the jth column as $\mathbf{e}_j^*$. If $\hat{\boldsymbol{\mu}}$ is the element of $\left\{\mathbf{Z}\hat{\boldsymbol{\beta}} : \hat{\boldsymbol{\beta}} \in \mathbb{R}^m\right\}$ that minimizes $(\mathbf{Y} - \hat{\boldsymbol{\mu}})^\intercal(\mathbf{Y} - \hat{\boldsymbol{\mu}})$, then,

$$\hat{\boldsymbol{\mu}} = \overline{\mathbf{Y}}_{1,2^q}\mathbf{e}_1^* + \overline{\mathbf{Y}}_{2^q+1,2\cdot 2^q}\mathbf{e}_2^* + \cdots + \overline{\mathbf{Y}}_{(m-1)2^q+1,m2^q}\mathbf{e}_m^*.$$

Thus, if the columns of $\mathbf{Z}_S$ spans $\text{span}(\mathbf{e}_1^*, \ldots, \mathbf{e}_m^*)$, then we must have that $\mathbf{Z}_S\hat{\boldsymbol{\beta}}_S = \hat{\boldsymbol{\mu}}$, too.

Let us choose $\mathbf{Z}_S = (\mathbf{F}_m)_{(2^q)}$ using the notation in Lemma 2.2 and Lemma 2.3. $\mathbf{F}_m$ is invertible, so it spans $\text{span}(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m)$. Copying the rows 2^q times, $(\mathbf{F}_{m+1})_{(2^q)}$ spans $\text{span}(\mathbf{e}_1^*, \ldots, \mathbf{e}_m^*)$. Define

$$\overline{\mathbf{Y}}^* = \begin{pmatrix} \overline{\mathbf{Y}}_{1,2^q} \\ \overline{\mathbf{Y}}_{2^q+1,2\cdot 2^q} \\ \vdots \\ \overline{\mathbf{Y}}_{(m-1)2^q+1,m2^q} \end{pmatrix}. \tag{2.7}$$

Since $\mathbf{Z}_S\hat{\boldsymbol{\beta}}_S = \hat{\boldsymbol{\mu}}$, if we remove duplicate rows,

$$\mathbf{F}_m\hat{\boldsymbol{\beta}}_S = \overline{\mathbf{Y}}^* \Rightarrow \hat{\boldsymbol{\beta}}_S = \mathbf{F}_m^{-1}\overline{\mathbf{Y}}^*, \tag{2.8}$$

which by Lemma 2.3 gives us

$$\begin{aligned} \hat{\beta}_{\mathbf{X}}^S &= F(m-1)\overline{\mathbf{Y}}_m^* + F(m-2)\overline{\mathbf{Y}}_{m-1}^* - \sum_{j=1}^{m-2} F(j)\overline{\mathbf{Y}}_j^* \\ &= F(m-1)\left(\overline{\mathbf{Y}}_m^* - \overline{\mathbf{Y}}_{m-1}^*\right) + \sum_{j=2}^{m-1} F(j+1)\left(\overline{\mathbf{Y}}_j^* - \overline{\mathbf{Y}}_{j-1}^*\right) + F(2)\overline{\mathbf{Y}}_1^*. \end{aligned} \tag{2.9}$$

Now, note that $\overline{\mathbf{Y}}_k^*$ is the mean of numbers taken between the $k-1$th m-quantile and kth

m-quantile, so

$$\begin{aligned}\mathbb{E}\,\overline{\mathbf{Y}}^*_k &= m\int_{\Phi^{-1}((k-1)/m)}^{\Phi^{-1}(k/m)} \frac{y}{\sqrt{2\pi}}\exp(-y^2/2)\,dy \\ &= \frac{m}{\sqrt{2\pi}}\left(\exp\left(-\frac{1}{2}\left[\Phi^{-1}\left(\frac{k-1}{m}\right)\right]^2\right) - \exp\left(-\frac{1}{2}\left[\Phi^{-1}\left(\frac{k}{m}\right)\right]^2\right)\right). \qquad (2.10)\end{aligned}$$

From Winitzki (2008), we have that

$$\begin{aligned}\Phi^{-1}(p) &= \sqrt{2}\,\mathrm{erf}^{-1}(2p-1) \qquad (2.11)\\ &\approx \sqrt{2}\,\mathrm{sgn}(2p-1)\sqrt{\sqrt{\left(\frac{2}{\pi a}+\frac{\log(4p-4p^2)}{2}\right)^2 - \frac{\log(4p-4p^2)}{a}} - \left(\frac{2}{\pi a}+\frac{\log(4p-4p^2)}{2}\right)},\end{aligned}$$

where $a = \dfrac{8(\pi-3)}{3\pi(4-\pi)}$.

Now looking at Equation 2.9 along with Equations 2.10 and 2.11, we see that $F(2)\overline{\mathbf{Y}}^*_1$ is the only negative term, and $F(m-1)\left(\overline{\mathbf{Y}}^*_m - \overline{\mathbf{Y}}^*_{m-1}\right) \xrightarrow{p} \infty$ as $m \to \infty$ since $F(m-1)$ increases exponentially.

Now, consider our test statistic

$$\mathbf{T} = \frac{\hat{\beta}^S_{\mathbf{X}}}{s\sqrt{\left((\mathbf{Z}_S^{\intercal}\mathbf{Z}_S)^{-1}\right)_{mm}}}. \qquad (2.12)$$

In the denominator, we have that

$$s^2 = \frac{1}{2N-m}\left(\mathbf{Y}-\mathbf{Z}_S\hat{\boldsymbol{\beta}}_S\right)^{\intercal}\left(\mathbf{Y}-\mathbf{Z}_S\hat{\boldsymbol{\beta}}_S\right) \le \frac{1}{2N-m}\sum_{i=1}^{2N} Y_i^2 \sim \frac{1}{2N-m}\chi^2_{2N}. \qquad (2.13)$$

Moreover, by Lemmas 2.2 and Lemma 2.3,

$$\begin{aligned}\left((\mathbf{Z}_S^{\intercal}\mathbf{Z}_S)^{-1}\right)_{mm} &= 2^{-q}\left(\sum_{j=1}^{m-1} F(j)^2 + F(m-2)^2\right) = 2^{-q}\left(F(m-1)F(m) + F(m-2)^2\right) \\ &= 2^{-q}\left(F(m-1)^2 + F(m-1)F(m-2) + F(m-2)^2\right) \\ &= 2^{-q}\left(F(m-1)^2 + F(m-1)(F(m-1)-F(m-3)) + F(m-2)^2\right) \\ &= 2^{-q}\left(2F(m-1)^2 + F(m-2)^2 - F(m-1)F(m-3)\right) \\ &= 2^{-q}\left(2F(m-1)^2 + (-1)^{m-1}\right), \qquad (2.14)\end{aligned}$$

where the last line follows by the Catalan identities.

Fix m, the number of coefficients and groups. Using Equation 2.10, we can bound $\overline{\mathbf{Y}}^*_m - \overline{\mathbf{Y}}^*_{m-1}$ below in probability as $N \to \infty$ by y_*. As $N \to \infty$, $q \to \infty$ if we fix m. Also, using Equation 2.13, we can bound s above in probability by s^*. Then, combining Equations 2.9, 2.12, and 2.14, we have that

$$\mathbb{P}\left(\mathbf{T} \geq 2^{-(q-1)/2}\frac{y_*}{s^*}\right) > 1 - \delta \tag{2.15}$$

for some $\delta > 0$.

So, for any $M > 0$ and $\epsilon > 0$, we can have that

$$\mathbb{P}(\hat{\beta}^S_{\mathbf{X}} > M) > 1 - \epsilon$$

by choosing large m by Equation 2.9. Once this m is chosen, we can fix m and choose large N to achieve statistical significance, that is,

$$\mathbb{P}\left(\mathbf{T} = \frac{\hat{\beta}^S_{\mathbf{X}}}{s\sqrt{\left(\left(\mathbf{Z}_S^\intercal \mathbf{Z}_S\right)^{-1}\right)_{mm}}} \geq T^{-1}_{2N-m}(1 - \alpha/2)\right) > 1 - \delta,$$

for any $\alpha > 0$ and $\delta > 0$ by Equation 2.15.

For the cases, where N is not divisible by m, we can have some groups of size $\lceil N/m \rceil$ and other groups of size $\lfloor N/m \rfloor$, and modify the definition of $\overline{\mathbf{Y}}^*$ accordingly. Equations 2.9 and 2.13 still hold, and we can derive analogs of Equations 2.10 and 2.14. □

The proof is instructive as it gives us an explicit matrix $\mathbf{Z}_S$. I have have verified the proof computationally with the results in Table 1.

3 Independent Y and X

In practice, we will not have an infinite number of covariates. To see how many are needed, I simulated collecting independent covariates and stopped when statistical significance was reached.

$2N$	m	Mean $\hat{\beta}^S_{\mathbf{X}}$	p-value below 0.05 (%)
64	32	1627097	1.2%
64	16	951.2453	39.3%
64	8	26.66189	99.7%
128	32	1752930	20.6%
128	16	991.1441	97.1%
128	8	27.33778	100%
256	64	7345141356974.77	10.5%
256	32	1833113	84.9%
256	16	1002.266	100%

Table 1: So indeed, we see that $\hat{\beta}^S_{\mathbf{X}}$ grows exponentially, and we can ensure statistical significance.

Consider the model in Equation 2.1 and define $\hat{\boldsymbol{\beta}}_S$ as in Equation 2.2. Now, suppose we followed best practices and did not perform any data dredging. That is, we specified S, and therefore $\mathbf{Z}_S$ before observing $\mathbf{Y}$. Our null hypothesis is $H_0 : \beta^S_{\mathbf{X}} = 0$ and the alternate hypothesis is $H_1 : \beta^S_{\mathbf{X}} \neq 0$.

By Chapter 6 of Bickel and Doksum (2015), $\hat{\beta}^S_{\mathbf{X}} \sim \mathcal{N}\left(0, \left((\mathbf{Z}_S^\intercal \mathbf{Z}_S)^{-1}\right)_{mm}\right)$ under the null hypothesis. Moreover, an unbiased esimator for σ^2 is $s^2 = \dfrac{|\mathbf{Y} - \mathbf{Z}_S\hat{\boldsymbol{\beta}}_S|^2}{2N - m}$, so

$$\frac{\hat{\beta}^S_{\mathbf{X}}}{s\sqrt{\left((\mathbf{Z}_S^\intercal \mathbf{Z}_S)^{-1}\right)_{mm}}} \sim \mathcal{T}_{2N-m}, \tag{3.1}$$

that is, the t distribution with $2N - m$ degrees of freedom. Thus, doing a two-sided t-test, the p-value associated with a set S is

$$p_S = 2\left(1 - T_{2N-m}\left(\left|\frac{\hat{\beta}^S_{\mathbf{X}}}{s\sqrt{\left((\mathbf{Z}_S^\intercal \mathbf{Z}_S)^{-1}\right)_{mm}}}\right|\right)\right), \tag{3.2}$$

where T_n is the cumulative distribution function of the t distribution with n degrees of freedom. For a test at level of significance α (usually 0.05), we reject the null hypothesis when $p_S \leq \alpha$.

Let $\mathbf{Y} \sim \mathcal{N}(0, I)$ and $\mathbf{X}$ be defined as in Equation 2.3, so $\mathbf{X}$ is a vector of N 0s and N 1s. $\mathbf{Y}$ and $\mathbf{X}$ are independent, so the real value of the coefficient is $\beta^S_{\mathbf{X}} = 0$. We simulate covariates $\{\mathbf{z}_j = (z_{1,j}, \ldots, z_{2N,j}) : j = 1, 2, \ldots\}$ such that $z_{i,j} \sim \text{Bernoulli}(1/2)$. Thus, the covariates are independent of $\mathbf{Y}$ and $\mathbf{X}$, so $\beta_j^S = 0$ for $j = 1, \ldots, m-1$. Since the null hypothesis is true, we would expect to reject the null hypothesis with probability α. More concretely, we make a Type I error and say that the treatment had some effect with probability α.

Now, one could imagine that a researcher wants the experiment to favor the treatment, so he or she collects a lot of data and chooses S after the experiment. By cheating, the researcher increases the probability of finding a statistically significant result or making a Type I error, depending on your perspective. By trying many different S, the researcher is testing multiple hypotheses and will eventually achieve statistical significance by chance. Once statistical significance is achieved, the researcher can pretend that he or she followed best practices. Just how easy is it to do this?

It was not possible for me to test every subset of covariates, so I used dynamic programming to choose S. Therefore, the simulations here give a conservative upper bound, and it may be possible to achieve statistical signficance with a smaller number of covariates.

Initialize $S_0 = \emptyset$. Upon drawing $\mathbf{z}_k$, where $k \in \mathbb{N}$, I defined sets $S_1, \ldots, S_k$ as

$$S_j = \begin{cases} S_{j-1} \cup \{\mathbf{z}_k\}, & j = k \\ S_{j-1} \cup \{\mathbf{z}_k\}, & p_{S_{j-1} \cup \{\mathbf{z}_k\}} < p_{S_j} \\ S_j, & \text{otherwise.} \end{cases} \tag{3.3}$$

The algorithm stops at k when there exists some j such that $p_{S_j} \leq 0.05$ at which point, we record the minimum such j, the number of covariates in the subset, and k, the number of total covariates drawn. Set $S = S_j$. Throughout the paper the k at which we stopped at will be referred to as the *set size*. The minimum j such that $p_{S_j} \leq 0.05$ will be referred to as the *subset size.*

Here are the results. For each N, 1000 simulations were run. Firstly, in a few cases, statistical significance was not reached after drawing $2N - 2$ linearly independent covariates when N was small as seen in Table 2. This problem disappears for larger N.

$2N$	Percent Significant
50	98.2%
100	100%
200	100%
400	100%
800	100%

Table 2: Statistical significance is almost always found before we have as many independent covariates as observations.

As we would expect for a level $\alpha = 0.05$ test, about 5% of the time, we found statistical significance immediately with 0 additional covariates. For the cases where more covariates are needed, in Table 3, we list the average set size needed (k in the algorithm) and the average subset size (j in the algorithm). As one can see, as N increases both averages increase but not by much every time we double N. Also, the distribution of the subset size is skewed right, and a subset size of 1 is most common. The percentage of such cases is listed in the third column.

$2N$	Mean Subset Size	Subset Size 1 (%)	Mean Set Size	SD Set Size
50	9.316	17.3%	22.935	12.515
100	12.107	15.6%	33.133	19.612
200	15.080	15.2%	44.168	29.188
400	16.774	16.1%	55.444	43.322
800	17.773	15.2%	70.127	59.416

Table 3: As N gets larger, we need bigger sets, but not that much bigger. Statistics were calculated after removing the cases of set size 0. Set Size refers to how many total covariates where drawn, that is, $|\{\mathbf{z}_j\}|$. Subset Size is the number of covariates actually used, that is, $|S|$.

The fourth column shows that not many covariates are needed compared to the number of observations, so the researcher does not have to search too hard for his or her covariates. Given a few dozen covariates, one can often reject the null hypothesis by only choosing single one, which makes the argument that best practices were followed quite defensible.

4 Power Boosting

In addition to the case when $\mathbf{Y}$ and $\mathbf{X}$ are independent, another situation where a researcher may want to dredge data to achieve statistical significance is when the effect of the treatment on $\mathbf{Y}$ is small and hard to detect, that is, the statistical power is small. The legitimate way to increase statistical power would be to collect more data, design a better experiment, or measure observations more precisely to decrease variance. For a variety of reasons such as budget or a deadline to publish, such corrections may not be possible, so the researcher may resort to data dredging.

First, pretend that we do the proper thing, follow best practices, and fix $S = \emptyset$, so we consider the model

$$\mathbf{Y} = \beta_{\mathbf{X}}^{S}\mathbf{X} + \boldsymbol{\epsilon}, \tag{4.1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)$. Consider the simple hypotheses, $H_0 : \beta_{\mathbf{X}}^{S} = 0$ versus $H_1 : \beta_{\mathbf{X}}^{S} = 1$. Suppose that σ^2 is known. The statistical power, $1 - \beta$, is the probability of correctly rejecting H_0 when H_1 is true, so $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}, \sigma^2 I)$, where β is the probability of a Type II error. As σ^2 increases, statistical power goes down.

Button et al. (2013) discusses many problems with low power studies including low reproducibility and overestimation of effect size. In order to boost the statistical power, a researcher may try to include other covariates. In order to simulate this data dredging, the first order of business is to establish how to choose σ^2 in order to achieve certain power.

4.1 Choosing σ^2

The main reference here is Chapter 6 of Bickel and Doksum (2015). Consider the model in Equation 4.1 with $2N$ observations where we have fixed $S = \emptyset$. Use the simple hypotheses

$$H_0 : \beta_{\mathbf{X}}^{S} = 0 \text{ and } H_1 : \beta_{\mathbf{X}}^{S} = 1.$$

Let $\mathbf{X}$ be defined as in Equation 2.3, where we assign N subjects to the treatment.

Then, $\mathbf{Y} \sim \mathcal{N}\left(\beta_{\mathbf{X}}^{S}\mathbf{X}, \sigma^2 I\right)$. By Corollary 6.1.1, $\hat{\boldsymbol{\beta}}_S \sim \mathcal{N}\left(\boldsymbol{\beta}_S, \sigma^2(\mathbf{Z}_S^{\intercal}\mathbf{Z}_S)^{-1}\right)$. In this case, $\hat{\boldsymbol{\beta}}_S = \hat{\beta}_{\mathbf{X}}^{S}$ and $\mathbf{Z}_S = \mathbf{X}$. Assume that σ^2 is known.

$2N$	Power $(1-\beta)$	Variance (σ^2)
50	0.100	58.7446
50	0.300	12.1477
50	0.500	6.50868
50	0.700	4.05055
100	0.100	117.489
100	0.300	24.2954
100	0.500	13.0174
100	0.700	8.1011
200	0.100	234.978
200	0.300	48.5908
200	0.500	26.0347
200	0.700	16.2022
400	0.100	469.957
400	0.300	97.1815
400	0.500	52.0695
400	0.700	32.4044
800	0.100	939.913
800	0.300	194.363
800	0.500	104.139
800	0.700	64.8088

Table 4: Higher variance means lower power.

So, under the null hypothesis,

$$\sqrt{N}\frac{\hat{\beta}_{\mathbf{X}}^{S}}{\sigma} \sim \mathcal{N}(0,1). \tag{4.2}$$

Thus, for a two-sided level α test, we will reject the null hypothesis if

$$\left|\sqrt{N}\frac{\hat{\beta}_{\mathbf{X}}^{S}}{\sigma}\right| \geq \Phi^{-1}\left(1-\frac{\alpha}{2}\right) = z_{\alpha/2}, \tag{4.3}$$

where Φ is the cumulative distribution function for the standard normal distribution.

To calculate statistical power, we assume that H_1 is true, so that actually

$$\sqrt{N}\frac{\hat{\beta}^S_{\mathbf{X}} - 1}{\sigma} \sim \mathcal{N}(0,1) \tag{4.4}$$

and calculate the probability of correctly rejecting H_0.

Suppose that we want to reject the null hypothesis with probability $1-\beta$. We apply Equation 4.3 and solve for σ^2:

$$\begin{aligned}
1 - \beta &= \mathbb{P}\left(\sqrt{N}\frac{\hat{\boldsymbol{\beta}}^S_{\mathbf{X}}}{\sigma} \geq z_{\alpha/2}\right) + \mathbb{P}\left(\sqrt{N}\frac{\hat{\boldsymbol{\beta}}^S_{\mathbf{X}}}{\sigma} \leq -z_{\alpha/2}\right) \\
&= \mathbb{P}\left(\sqrt{N}\frac{\hat{\beta}^S_{\mathbf{X}} - 1}{\sigma} \geq z_{\alpha/2} - \frac{\sqrt{N}}{\sigma}\right) + \mathbb{P}\left(\sqrt{N}\frac{\hat{\beta}^S_{\mathbf{X}} - 1}{\sigma} \leq -z_{\alpha/2} - \frac{\sqrt{N}}{\sigma}\right) \\
&= \left[1 - \Phi\left(z_{\alpha/2} - \frac{\sqrt{N}}{\sigma}\right)\right] + \Phi\left(-z_{\alpha/2} - \frac{\sqrt{N}}{\sigma}\right) \\
&= \Phi\left(-z_{\alpha/2} + \frac{\sqrt{N}}{\sigma}\right) + \Phi\left(-z_{\alpha/2} - \frac{\sqrt{N}}{\sigma}\right).
\end{aligned}$$

This can be accomplished with a binary search. The computed variances can be found in Table 4.

4.2 Simulation

As one can see from Table 4, with high variances, the power is rather low, so despite the treatment having an effect on $\mathbf{Y}$, the probability of finding statistical significance can be very low. The researcher may find this situation unacceptable, not be able to increase statistical power through legitimate means, and thus, feel the need to cheat by including other covariates. Given a nearly statistically significant result, one may avail oneself of illegitimate means to decrease the p-value such as varying S. Again, we try to answer the question of how easy is it to construct such S.

Simulations were run at 4 power levels: 0.1, 0.3, 0.5, and 0.7. The results can be found in Table 5. The direction of the results are largely expected. As the power gets larger, fewer covariates and a smaller subset are needed to construct S. While the direction is not surprising, perhaps the small

$2N$	Power (1 - β)	Mean Subset Size	Subset Size 1 (%)	Mean Set Size	SD Set Size
50	0.050	9.316	17.3%	22.935	12.515
50	0.100	10.161	18.5%	22.743	12.889
50	0.300	8.720	30.4%	19.846	13.749
50	0.500	7.855	36.5%	18.488	13.592
50	0.700	5.879	47.6%	16.726	13.113
100	0.050	12.107	15.6%	33.133	19.612
100	0.100	12.342	17.9%	31.059	20.217
100	0.300	9.557	31.9%	24.863	19.780
100	0.500	8.374	41.9%	22.458	20.159
100	0.700	7.743	48.2%	21.743	20.157
200	0.050	15.080	15.2%	44.168	29.188
200	0.100	14.966	21.3%	41.698	30.197
200	0.300	11.343	32.3%	34.343	29.155
200	0.500	9.980	43.8%	31.262	30.123
200	0.700	6.926	52.8%	23.980	25.789
400	0.050	16.774	16.1%	55.444	43.322
400	0.100	16.570	19.7%	53.071	42.845
400	0.300	14.025	33.5%	43.711	42.415
400	0.500	10.263	44.7%	34.586	37.821
400	0.700	9.312	54%	32.130	40.195
800	0.050	17.773	15.2%	70.127	59.416
800	0.100	18.948	20.6%	66.160	59.417
800	0.300	12.558	34.7%	51.478	54.661
800	0.500	9.725	46.7%	40.853	51.674
800	0.700	7.062	56.5%	31.455	44.540

Table 5: As the power increases, we need less covariates, and the subset size decreases, too. Power 0.05 corresponds to the case where $\mathbf{Y}$ and $\mathbf{X}$ are independent.

size of our set of covariates and subset is. Even at the very low power of 0.3, $|S| = 1$ one-third of the time. At power 0.7, about half the time only 1 covariate is needed in the subset. Moreover, the total amount of additional covariates that need to be collected (Set Size) is small relative to N.

If the test is underpowered, cheating with a supply of independent covariates, one effectively increases the power. In many fields of research, a power of 0.8 is standard McDonald (2009). For this reason, the minimum set size such that the null hypothesis was rejected in 80% of cases is of interest. This threshold can be seen in the third column of Table 6. We find that while we need a substantial number of covariates to achieve this at low power, the the size of our subset is rather small. When statistical power is 0.5, the number of total additional covariates needed was less than 30. Out of the covariates collected, often less than 10 were needed to construct S. These numbers barely changed with the number of observations. So, if one allows this sort of cheating, an experiment that is reproducible 50% of time becomes reproducible 80% of the time by adding a handful of cherry-picked covariates.

$2N$	Power (1 - β)	80% Set Size	80% Subset Size	Mean Subset Size (80%)
50	0.050	35	34	6.912
50	0.100	34	34	7.103
50	0.300	30	29	4.872
50	0.500	22	16	2.584
50	0.700	7	2	1.149
100	0.050	50	39	8.775
100	0.100	49	41	7.850
100	0.300	38	27	4.533
100	0.500	26	13	1.839
100	0.700	7	3	1.135
200	0.050	72	49	9.811
200	0.100	69	53	8.720
200	0.300	55	41	4.385
200	0.500	30	11	1.577
200	0.700	6	3	1.137
400	0.050	99	69	8.632
400	0.100	95	63	8.096
400	0.300	72	36	2.996
400	0.500	27	6	1.500
400	0.700	4	2	1.099
800	0.050	134	85	6.327
800	0.100	129	74	6.353
800	0.300	74	36	2.346
800	0.500	26	5	1.450
800	0.700	5	2	1.088

Table 6: If we only want an 80% power test, the set and subset sizes become much smaller.

4.2.1 Effect Size

$2N$	Power (1 - β)	Percent $\hat{\beta}^S_{\mathbf{X}} > 0$	Mean $\hat{\beta}^S_{\mathbf{X}}$ when $\hat{\beta}^S_{\mathbf{X}} > 0$
50	0.100	69.8%	4.480
50	0.300	84.9%	1.923
50	0.500	93.7%	1.431
50	0.700	97.4%	1.168
100	0.100	66.7%	4.186
100	0.300	86.7%	1.873
100	0.500	92.2%	1.386
100	0.700	96.7%	1.173
200	0.100	67.2%	4.050
200	0.300	85.8%	1.818
200	0.500	94.4%	1.361
200	0.700	97.6%	1.163
400	0.100	68.1%	3.983
400	0.300	87.3%	1.810
400	0.500	93.5%	1.370
400	0.700	97%	1.142
800	0.100	70.9%	3.948
800	0.300	88.4%	1.793
800	0.500	92.9%	1.361
800	0.700	97.7%	1.152

Table 7: Even at low power, we find the correct effect despite data dredging. However, the size of the effect is overestimated at these low powers.

The actual value is $\beta^S_{\mathbf{X}} = 1$. At best, one would hope that $\hat{\beta}^S_{\mathbf{X}}$ is close to 1, and at the very least, one would hope that $\hat{\beta}^S_{\mathbf{X}} > 0$ so the direction of the effect is correct. In Table 7, we see that even at low power, the direction of the effect is usually correct.

Define experimental effect size as the magnitude of $\hat{\beta}^S_{\mathbf{X}}$. While the direction may be correct, when the power is 0.1, the experimental effect size is nearly 4 times the actual effect size. When the power is 0.3, the experimental effect size is about 80% larger. When the power is 0.5, the experimental effect size is about 40% larger. And when the power is 0.7, the experimental effect size is about 15% larger. So, the results agree with Button et al. (2013) and support the findings of Open Science

Collaboration (2015), which found that upon reproducing the studies, mean effect size was only half as large as the original study.

5 Discussion

It should now be clear that one can easily manipulate one's data by data dredging and doing multiple hypothesis testing. In the first part, I showed mathematically that if you collect enough data irrelevant to your independent and dependent variables, you can not only find statistical significance but significantly overstate your effect size. The second section discusses how easily one can find an effect that does not exist. Much of the time this can be done with a single covariate, so a researcher could obscure the fact that multiple hypothesis testing was done. Finally, the third part shows that when the effect does exist, data dredging leads to a false sense of reproducibility and misidentifying the effect either in direction or magnitude. In this manner, poorly designed studies may vastly overstate the importance of their findings.

While it was alreadly well-known that data can be manipulated to find statistical significance, this work reveals exactly how easily it can be done in the case of a balanced treatment and classical linear regression. The ease of computability allows one to exploit the closed-form solutions to increase the effect size without bound, and yet, maintain statistical significance. In the simulations, while a large number of covariates had to be generated at times, the actual subset needed was usually very small. Even in cases with a large amount of observations, often one or two covariates sufficed. These small subsets lend a false sense of legitimacy to these models obtained by multiple hypothesis testing. It would not be hard for a dishonest researcher to claim that he or she specified the ill-gotten model beforehand after mining a small amount of data.

For honest researchers, these results emphasize the importance of adhering to the prescriptions of

Senn (1994): (1) randomly assigning the treatment, (2) identify covariates of prognostic value before the experiment and including all of them in the regression, and (3) only look at additional covariates post-study to inform future models and experiments. Moreover, statistical power should be increased in legitimate ways by increasing the sample size, better study design, or more precise measurement. Reproduction of experiments must be done to ensure validity of the results and magnitude of the effect size.

It remains to establish the results of the simulations theoretically, that is, given a large number of independent covariates, only a small number are required in the regression to get statistical significance. Morever, the results in this paper only deal with indicator variables. These results would be expected to hold for covariates with other distributions such as standard normal.

All in all, this paper gives credence to Ioannidis' assertion that most research findings are false. In particular, when researchers are not transparent about their experimental design and all the hypotheses tested, they could easily generate credible models through data dredging. Another implication is that many statistically significant discoveries may not be all that important due to manipulation of the effect size. Therefore, these results indicate a greater need for data transparency and reproducibility.

6 Code

The C++ code for the simulations can be found on GitHub. [1] Boost (Boost, 2002) was used for random number generation. Armadillo was used for linear algebra routines (Sanderson, 2010). Data analysis was done in R with the `data.table` package (R Core Team, 2015; Dowle et al., 2014). Tables were generated with the `xtable` package (Dahl, 2016).

[1] `https://github.com/ppham27/cheating-linear-models-simulations`

7 Acknowledgements

This thesis was written under the supervision of Professor Robin Pemantle. I would like to thank him for his guidance and advice. I would also like to extend my gratitude towards Professors Charles Epstein and Dylan Small for serving on my committee.

References

Bickel, P. and K. Doksum (2015). *Mathematical Statistics: Basic Ideas and Selected Topics, Volume I, Second Edition.* Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

(2002). *The Boost Graph Library: User Guide and Reference Manual.* Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Button, K. S., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafo (2013, May). Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci 14*(5), 365–376. Analysis.

Dahl, D. B. (2016). *xtable: Export Tables to LaTeX or HTML.* R package version 1.8-2.

Dowle, M., T. Short, S. Lianoglou, A. S. with contributions from R Saporta, and E. Antonyan (2014). *data.table: Extension of data.frame.* R package version 1.9.4.

Durrett, R. (2010). *Probability: Theory and Examples.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Gelman, A. and K. O'Rourke (2014). Discussion: Difficulties in making inferences about scientific truth from distributions of published p-values. *Biostatistics 15*(1), 18–23.

Gilbert, D. T., G. King, S. Pettigrew, and T. D. Wilson (2016). Comment on "estimating the reproducibility of psychological science". *Science 351*(6277), 1037–1037.

Ioannidis, J. P. A. (2005, 08). Why most published research findings are false. *PLoS Med 2*(8).

Ioannidis, J. P. A. (2014). Discussion: Why an estimate of the science-wise false discovery rate and application to the top medical literature is false. *Biostatistics 15*(1), 28–36.

Jager, L. R. and J. T. Leek (2014). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics 15*(1), 1–12.

McDonald, J. H. (2009). *Handbook of biological statistics*, Volume 2. Sparky House Publishing Baltimore, MD.

Moonesinghe, R., M. J. Khoury, and A. C. J. W. Janssens (2007, 02). Most published research findings are false—but a little replication goes a long way. *PLoS Med 4*(2), 1–4.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science 349*(6251).

Permutt, T. (1990). Testing for imbalance of covariates in controlled experiments. *Statistics in Medicine 9*(12), 1455–1462.

R Core Team (2015). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Sanderson, C. (2010, oct). Armadillo: An open source c++ linear algebra library for fast prototyping and computationally intensive experiments. In *NICTA*, Australia.

Senn, S. (1994). Testing for baseline balance in clinical trials. *Statistics in Medicine 13*(17), 1715–1726.

Winitzki, S. (2008). A handy approximation for the error function and its inverse. *A lecture note obtained through private communication.*