Estimates on the boundary in random regressions:

Statistical methods research done as science rather than math

(Work in progress)
This lecture is about two things.

- The manner in which we statistical academics study statistical methods.
- A problem that commonly occurs in using a particular model to analyze data.

I’ll begin by describing the latter, which leads to the former.
The random regressions model

Data are grouped in clusters $i; j$ indexes observations within clusters.

The outcome $y_{ij}$ is presumed to arise as

$$y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \epsilon_{ij}, i = 1, \ldots, N, j = 1, \ldots, s,$$  \hspace{1cm} (1)

For today, $x_{ij}$ is scalar

$\epsilon_{ij}$ are iid $N(0, \sigma^2_e)$

$(\beta_{0i}, \beta_{1i})'$ are iid $N((b_0, b_1)', \Sigma)$, with

$$\Sigma = \begin{bmatrix} \sigma^2_c & \rho \sigma_c \sigma_s \\ \rho \sigma_c \sigma_s & \sigma^2_s \end{bmatrix}$$  \hspace{1cm} (2)

Subscripts "c", "s" are for intercepts $\beta_{0i}$, slopes $\beta_{1i}$ respectively.
The RL-maximizing estimates \((\hat{\sigma}_c^2, \hat{\sigma}_s^2, \hat{\rho})\) can be on the boundary of legal values:

\[
\hat{\rho} = -1 \text{ or } +1, \text{ or } \hat{\sigma}_c^2 = 0, \text{ or } \hat{\sigma}_s^2 = 0
\]

In my experience this happens often enough to be a real problem.

This lecture is mostly about \(\hat{\rho} = \pm 1\).

Our software gives us useless or misleading information when this happens because our theory says little or nothing about this subject.
Here’s what we have instead of actual knowledge

I googled “lme4 and nlme give different results”

First hit (dated 2 October 2012): “I got completely different results from lmer() and lme()! ... Why is that and which model is correct?”

The result from lme looks reasonable enough, but ...

```r
> mix1d = lme(logInd ~ 0 + crit_i + Year:crit_i, random = ~ 1 + Year | Taxon, data = datai)
> summary(mix1d)
Linear mixed-effects model fit by REML
Data: datai
   AIC      BIC   logLik
4727.606 4799.598 -2351.803

Random effects:
Formula: ~1 + Year | Taxon
Structure: General positive-definite, Log-Cholesky parametrization
       StdDev      Corr
(Intercept) 9.829727e-08 (Intr)
   Year       3.248182e-04  0.619
Residual    4.933979e-01
```
In the result from lmer, the estimated correlation $\hat{\rho}$ is -1.
Laplace approximation is used with non-normal errors ⇒ irrelevant.

15 fake datasets where lmer’s \( \hat{\rho} = -1 \) and lmer’s didn’t: lmer found the global max in all, lme didn’t even find a local maximum (except once).

lme was fooled by a nearly flat RL.
Why are we dumb like this? How do we change that?

We generally use math to try to understand our methods.

If they’re too complicated for that, we formulate a different math problem, usually a large-sample problem, and solve that instead.

Unfortunately, except for suggesting approximations, asymptotics provide no practically useful information and can be misleading.

Large-sample theorems usually begin by assuming the true parameter value is not on the boundary, so if the sample is large enough then the estimate is not on the boundary. (Proof by assumption.)
There is an alternative

Instead of producing useless information and sometimes fooling ourselves, we can and should mimic what biologists do to understand Nature’s black boxes:

▶ Capture the phenomenon of interest in a simple model system, not a statistical model but rather “a usually miniature representation of something”, e.g., an animal or cell-culture model.
▶ Hypothesize about how the phenomenon of interest works.
▶ Do experiments with the model system to test those hypotheses.
▶ Iterate; revise the model system and hypotheses as needed.

I’ll demonstrate this approach by asking, for the RR model, what conditions make it likely that \( \hat{\rho} = \pm 1 \).
The approach, for the random-regressions problem

Methods research question: Which features of the design or data-generating process influence the chance that $\hat{\rho} = \pm 1$?

In the biological style of investigation, the components are:

- The model system: $N$ clusters with the same design matrix, having 2 orthogonal columns. The following features of the model system can be varied: $N$, within-cluster sample size $s$, $\rho$, $\sigma_c^2$, $\sigma_s^2$, $\beta$.
- Generate hypotheses about the chance of getting $\hat{\rho} = \pm 1$ using a simplified “predictor”.
- Test those hypotheses using simulation experiments and accumulate facts that are either immediately useful or that need to be explained.
- Revise hypotheses as indicated.
The model system

\[ y_{ij} = \beta_{0i} + \beta_{1i}x_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, N, \quad j = 1, \ldots, s \]

\( \epsilon_{ij} \) are iid \( N(0, \sigma^2_e) \)

\( (\beta_{0i}, \beta_{1i})' \) are iid \( N((b_0, b_1)', \Sigma) \), \( \Sigma = \begin{bmatrix} \sigma^2_c & \rho \sigma_c \sigma_s \\ \rho \sigma_c \sigma_s & \sigma^2_s \end{bmatrix} \)

Cluster sample size \( s = 2m + 1 \) for \( m \) a positive integer.

Vector of predictors \( x_i = (-1, -(m - 1)/m, \ldots, 0, \ldots, (m - 1)/m, 1)' \).
In mixed linear model notation

\[ H = s \times 2 \text{ design matrix for a cluster, with columns } \mathbf{1}_s \text{ and } \mathbf{x}_i. \]

We have \( N \) clusters and \( n = Ns \) total observations, and

\[
\begin{align*}
\mathbf{X} &= \mathbf{1}_N \otimes H \quad \text{is } n \times 2 \\
\beta &= (b_0, b_1)' \quad \text{is } 2 \times 1 \\
\mathbf{Z} &= \mathbf{I}_N \otimes H \quad \text{is } n \times 2N \\
\mathbf{u} &= (u_{10}, u_{11}|u_{20}, u_{21}|\ldots|u_{N0}, u_{N1})' \quad \text{is } 2N \times 1 \\
\mathbf{G} &= \text{BlockDiag}(\Sigma) \quad \text{is } 2N \times 2N, \text{ and} \\
\mathbf{R} &= \sigma_e^2 \mathbf{I}_n.
\end{align*}
\]

\( \otimes \) is the Kronecker product, \( \mathbf{A} \otimes \mathbf{B} = (a_{ij} \mathbf{B}) \)
The RL is straightforward, with these assumptions.

It’s *relatively* simple, thanks to the assumptions, but still way too complicated to allow much intuition.

Thus, I’m not going to show it.

To develop intuition and hypotheses, I made some simplifying assumptions to produce something I’ll call a predictor of when $\hat{\rho} = -1$. 
Steps in deriving the predictor

A necessary condition for \( \hat{\rho} = -1 \) is:

\[
\frac{\partial}{\partial \rho} \log \text{RL or log profiled RL, evaluated at } \rho = -1, < 0.
\]

The predictor is a fairly gross approximation to this derivative.

The steps in developing the predictor are:
- simplify the log RL,
- profile out one unknown,
- get \( \frac{\partial}{\partial \rho} \) of the profiled simplified log RL at \( \rho = -1 \),
- make more simplifications.

We have some freedom here because our goal is merely to create a device for generating hypotheses.
Simplify: Let $\sigma_c^2 = \sigma_s^2 \equiv \sigma_r^2$.

Profile out $\sigma_r^2$. The profiled log RL is a function of $\rho$ and $r = \sigma_e^2/\sigma_r^2$.

Take $\partial/\partial \rho$ of the profiled simplified log RL at $\rho = -1$.

Simplify more: Replace functions of the data by their expected values.

The predictor for $\hat{\rho} = -1$ (the predictor for $\hat{\rho} = +1$ is very similar):

$$
\left(\frac{N_s - N - 1}{(1 + r/s)(1 + r/q)}\right) \left[1 - \left(\frac{N_s - 2}{N_s - N - 1}\right) \frac{1 - \frac{N-1}{N(s-2)} \rho}{1 + \frac{2(N-1)(1+r/s)(1+r/q)+\rho}{N(s-2)(1+r/s)(1+r/q)-1}}\right].
$$

Note $s = 2m + 1$, $q = (2m^2 + 3m + 1)/3m$ are about cluster size.

Easy to show: predictor $> 0$ for all legal $N$, $s$, $r$, and $\rho \in (-1, 1)$. 
Generating hypotheses

The predictor; \( r = \frac{\sigma_e^2}{\sigma_r^2} \), \( s = 2m + 1 \), \( q = \frac{(2m^2 + 3m + 1)}{3m} \).

\[
\left( \frac{Ns - N - 1}{(1 + r/s)(1 + r/q)} \right) \left[ 1 - \left( \frac{Ns - 2}{Ns - N - 1} \right) \right] \frac{1 - \frac{N-1}{N(s-2)} \rho}{1 + \frac{2(N-1)}{N(s-2)} \frac{(1+r/s)(1+r/q)+\rho}{(1+r/s)(1+r/q)-1}}.
\]

Easy to prove mathematically:

- Given \( N \), \( s \), and \( \rho \), as \( r \) increases the predictor \( \to 0 \).
- Given \( \rho \) and \( r \), as \( N \) or \( s \) increases, the predictor \( \to \infty \).
- Given \( N \), \( s \), and \( r \), as \( \rho \) goes to \(-1\), the predictor \( \to 0 \).

Easy to show with simulations: small predictor \( \Rightarrow \) high chance \( \hat{\rho} = -1 \)

large predictor \( \Rightarrow \) small chance \( \hat{\rho} = -1 \).

\( \Rightarrow \) We have three hypotheses about \( \hat{\rho} \)’s gross behavior.
Interpreting the hypotheses

If these hypotheses are correct

\[ \hat{\rho} = -1 \] is more likely as \( \sigma_e^2 \) increases relative to \( \sigma_c^2, \sigma_s^2 \)
less likely as \( N \) or \( s \) increases.

\[ \Rightarrow \hat{\rho} = -1 \] mainly arises from poor resolution in the study design.

More interesting: if the study design’s resolution is poor enough,
\[ \hat{\rho} = +1 \] and \( \hat{\rho} = -1 \) are about equally likely for any true \( \rho \).
Developing more quantitative hypotheses

I drew 1000 sets of \((N, s, \rho, r)\):

\[
N \in \{50, 150, 250, \ldots, 1050\},
\]
\[
s \in \{5, 15, 25, \ldots, 105\},
\]
\[
\rho \in \{-0.9, -0.8, \ldots, -0.2, -0.1\},
\]
\[
\log_{10} r \in \{-2, -1.6, -1.2, \ldots, 2\}.
\]

I computed \(\log_{10}\) predictor and analyzed it using ANOVA.

Further hypotheses arising from this exercise:

- An increase in \(r\) can be countered by an increase in \(N\) or \(s\).
- Changes in \(s\) have more leverage in offsetting changes in \(r\) than do changes in \(N\). Multiplying or dividing \(r\) by about 2.5 is countered by multiplying or dividing \(N\) by about 5 or \(s\) by a factor of about 3.
- Some changes in \(r\) are so large that no change in \(\rho\) can un-do them.
Testing the hypotheses using simulation experiments

(1) To derive the predictor, I set $\sigma_c^2 = \sigma_s^2$.

In the experiments, I simulated data by setting $\sigma_c^2 = \sigma_s^2$.

I fit models allowing $\sigma_c^2$ and $\sigma_s^2$ to be different.

(2) Preliminary simulations identified predictor values giving some but not too many “bad” estimates.

A “bad” estimate is $\hat{\rho} = \pm 1$ or $\sigma_c^2 = 0$ or $\sigma_s^2 = 0$.

This is the region in which changes in $(N, s, \rho, r)$ can most affect the fraction of bad estimates $\Rightarrow$ most useful for testing the hypotheses.

(3) Datasets simulated from the RR model: $\beta = (0, 0)'$, $\sigma_c^2 = \sigma_s^2 = 1$.

(4) All analyses used the lmer function in the R package lme4.
A bit more about the experiments

Recall that $r = \frac{\sigma_e^2}{\sigma_c^2} = \frac{\sigma_c^2}{\sigma_s^2}$ in the simulated data (not the fit).

I did the experiments in three groups:

1. Examining how increasing $r$ produces a large fraction of bad estimates irrespective of $N$, $s$, or $\rho$.
2. Studying the tradeoffs between $r$ on the one hand and $N$, $s$, and $\rho$ on the other hand.
3. Learning more about the effect of $\rho$. 
Increasing $r$ produces many bad estimates

<table>
<thead>
<tr>
<th>Experiment A</th>
<th>predictor</th>
<th>% with $\hat{\rho}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>$s$</td>
<td>$\rho$</td>
</tr>
<tr>
<td>500</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>500</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>500</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>500</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>500</td>
<td>21</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment B</th>
<th>predictor</th>
<th>% with $\hat{\rho}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>$s$</td>
<td>$\rho$</td>
</tr>
<tr>
<td>500</td>
<td>21</td>
<td>0.95</td>
</tr>
<tr>
<td>500</td>
<td>21</td>
<td>0.95</td>
</tr>
<tr>
<td>500</td>
<td>21</td>
<td>0.95</td>
</tr>
<tr>
<td>500</td>
<td>21</td>
<td>0.95</td>
</tr>
<tr>
<td>500</td>
<td>21</td>
<td>0.95</td>
</tr>
</tbody>
</table>
Points re Experiments A, B – 100 datasets/setting

- Other experiments (below) support the same conclusion, especially Experiment G.
- \( N = 500 \) and \( s = 21 \) are quite a bit larger than in any dataset to which I have fit random regressions.
- Datasets giving \( \hat{\rho} = \text{NaN} \) had either \( \hat{\sigma}^2_c = 0 \) or \( \hat{\sigma}^2_s = 0 \); \( \hat{\sigma}^2_c = 0 \) in all cases I’ve examined.
- Given \( N \), \( s \), and \( r \), the chance of a “bad” estimate is minimized by setting \( \rho = 0 \), and Experiment A does this.
- BUT a true \( \rho \) close to +1 has no effect on \( \hat{\rho} = -1 \) for the largest \( r \).
- Oddity: With \( \rho = 0 \) and large \( r \), \( \hat{\rho} = +1 \) is more likely than \( \hat{\rho} = -1 \).
Tradeoffs of $r$ versus $N$, $s$, and $\rho$

Experiments C through F all have the same structure:

- Setting 1 is a base case.
- In setting 2, $r$ is changed.
- Settings 3, 4, 5 try to offset that by changing $N$, $s$, $\rho$ respectively.

For each Experiment:

- The base $r$ was chosen to give some “bad” estimates.
- For setting 2, $r$ was changed by a factor of $10^{0.4}$ compared to setting 1.
- In settings 3, 4, and 5, $N$, $s$, and $\rho$ were chosen to change the predictor of $\hat{\rho} = -1$ back to its original value.

400 datasets/setting except Experiment D, which had 600/setting.
## Experiments C and D: Moderate $N$ and $s$

### Experiment C

<table>
<thead>
<tr>
<th>Setting</th>
<th>$N$</th>
<th>$s$</th>
<th>$\rho$</th>
<th>$r$</th>
<th>Predictor $-1$</th>
<th>Predictor $+1$</th>
<th>% with $\hat{\rho}$ $-1$</th>
<th>$\hat{\rho}$ $+1$</th>
<th>NaN</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>9</td>
<td>-0.8</td>
<td>6.3</td>
<td>8.2</td>
<td>67</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>9</td>
<td>-0.8</td>
<td>15.8</td>
<td>1.7</td>
<td>15</td>
<td>39</td>
<td>0</td>
<td>2</td>
<td>41</td>
</tr>
<tr>
<td>3</td>
<td>500</td>
<td>9</td>
<td>-0.8</td>
<td>15.8</td>
<td>8.4</td>
<td>74</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>25</td>
<td>-0.8</td>
<td>15.8</td>
<td>8.7</td>
<td>76</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>9</td>
<td>0.0</td>
<td>15.8</td>
<td>8.2</td>
<td>8.2</td>
<td>8</td>
<td>6</td>
<td>17</td>
<td>30</td>
</tr>
</tbody>
</table>

### Experiment D

<table>
<thead>
<tr>
<th>Setting</th>
<th>$N$</th>
<th>$s$</th>
<th>$\rho$</th>
<th>$r$</th>
<th>Predictor $-1$</th>
<th>Predictor $+1$</th>
<th>% with $\hat{\rho}$ $-1$</th>
<th>$\hat{\rho}$ $+1$</th>
<th>NaN</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>9</td>
<td>-0.8</td>
<td>15.8</td>
<td>1.66</td>
<td>14.60</td>
<td>42</td>
<td>1</td>
<td>1</td>
<td>44</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>9</td>
<td>-0.8</td>
<td>6.3</td>
<td>8.23</td>
<td>67.47</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>9</td>
<td>-0.8</td>
<td>6.3</td>
<td>1.67</td>
<td>13.69</td>
<td>48</td>
<td>1</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>3</td>
<td>-0.8</td>
<td>6.3</td>
<td>1.83</td>
<td>15.14</td>
<td>42</td>
<td>0</td>
<td>0</td>
<td>42</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>9</td>
<td>-0.96</td>
<td>6.3</td>
<td>1.66</td>
<td>72.82</td>
<td>47</td>
<td>0</td>
<td>0</td>
<td>47</td>
</tr>
</tbody>
</table>
Experiments E and F: \( N, s \) have one large, one small

<table>
<thead>
<tr>
<th>Experiment E</th>
<th>( N )</th>
<th>( s )</th>
<th>( \rho )</th>
<th>( r )</th>
<th>predictor (-1)</th>
<th>predictor (+1)</th>
<th>( % ) with ( \hat{\rho} ) (-1)</th>
<th>( \hat{\rho} ) (+1)</th>
<th>NaN</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>25</td>
<td>-0.8</td>
<td>6</td>
<td>9.88</td>
<td>79.92</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>25</td>
<td>-0.8</td>
<td>15</td>
<td>1.85</td>
<td>16.02</td>
<td>42</td>
<td>2</td>
<td>2</td>
<td>45</td>
</tr>
<tr>
<td>3</td>
<td>104</td>
<td>25</td>
<td>-0.8</td>
<td>15</td>
<td>10.00</td>
<td>86.64</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>63</td>
<td>-0.8</td>
<td>15</td>
<td>9.66</td>
<td>83.30</td>
<td>23</td>
<td>0</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>25</td>
<td>0.0</td>
<td>15</td>
<td>9.06</td>
<td>9.06</td>
<td>9</td>
<td>10</td>
<td>4</td>
<td>23</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment F</th>
<th>( N )</th>
<th>( s )</th>
<th>( \rho )</th>
<th>( r )</th>
<th>predictor (-1)</th>
<th>predictor (+1)</th>
<th>( % ) with ( \hat{\rho} ) (-1)</th>
<th>( \hat{\rho} ) (+1)</th>
<th>NaN</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1000</td>
<td>3</td>
<td>-0.8</td>
<td>9</td>
<td>10.05</td>
<td>86.15</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>1000</td>
<td>3</td>
<td>-0.8</td>
<td>23</td>
<td>1.88</td>
<td>16.77</td>
<td>38</td>
<td>0</td>
<td>0</td>
<td>38</td>
</tr>
<tr>
<td>3</td>
<td>5350</td>
<td>3</td>
<td>-0.8</td>
<td>23</td>
<td>10.08</td>
<td>89.81</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>1000</td>
<td>9</td>
<td>-0.8</td>
<td>23</td>
<td>8.78</td>
<td>77.92</td>
<td>19</td>
<td>0</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>1000</td>
<td>3</td>
<td>0.0</td>
<td>23</td>
<td>9.36</td>
<td>9.36</td>
<td>10</td>
<td>6</td>
<td>0</td>
<td>16</td>
</tr>
</tbody>
</table>
Points re Experiments C, D, E, F

- I used $\rho = -0.8$ because $\rho < 0$ is far more plausible than $\rho > 0$.

- The chosen $N$ and $s$ offset the change in $r \approx$ as predicted.
  - Testing setting 3 vs. setting 1 gives large P except for Experiment E, where the increase in $N$ overcorrects.
  - Testing setting 4 vs. setting 1 gives large P in all four experiments.
  - Proportional changes offsetting the change in $r$:
    - $s$: 2.8, 3, 2.5, 3.
    - $N$: 5, 4.8, 5.2, 5.4.

- $\rho$ does not trade off against $r$ as the predictor suggested.
  - The predictor says a big change in $\rho$ offsets the change in $r$.
  - A big change in $\rho$ reduces the chance $\hat{\rho} = -1$, but also increases the chance $\hat{\rho} = +1$ or NaN.
Experiment G: More on the effect of $\rho$

All settings have the same (fairly large) $N$ and $s$.

Each block of 5 settings has one $r$ and $\rho$ ranging from $-0.95$ to $+0.95$.

400 datasets/setting

<table>
<thead>
<tr>
<th>setting</th>
<th>N</th>
<th>s</th>
<th>rho</th>
<th>r</th>
<th>predictor</th>
<th>% with $\hat{\rho}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-1</td>
<td>+1</td>
</tr>
<tr>
<td>1</td>
<td>500</td>
<td>21</td>
<td>-0.95</td>
<td>53</td>
<td>1.00</td>
<td>38.65</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>21</td>
<td>-0.50</td>
<td>53</td>
<td>9.96</td>
<td>29.78</td>
</tr>
<tr>
<td>3</td>
<td>500</td>
<td>21</td>
<td>0.00</td>
<td>53</td>
<td>19.89</td>
<td>19.89</td>
</tr>
<tr>
<td>4</td>
<td>500</td>
<td>21</td>
<td>0.50</td>
<td>53</td>
<td>29.78</td>
<td>9.96</td>
</tr>
<tr>
<td>5</td>
<td>500</td>
<td>21</td>
<td>0.95</td>
<td>53</td>
<td>38.65</td>
<td>1.00</td>
</tr>
</tbody>
</table>
But look what happens when $r$ increases

### Experiment G – larger $r$

<table>
<thead>
<tr>
<th>setting</th>
<th>N</th>
<th>s</th>
<th>rho</th>
<th>r</th>
<th>predictor</th>
<th>% with $\hat{\rho}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-1</td>
<td>+1</td>
</tr>
<tr>
<td>6</td>
<td>500</td>
<td>21</td>
<td>-0.95</td>
<td>271</td>
<td>0.05</td>
<td>1.94</td>
</tr>
<tr>
<td>7</td>
<td>500</td>
<td>21</td>
<td>-0.50</td>
<td>271</td>
<td>0.50</td>
<td>1.50</td>
</tr>
<tr>
<td>8</td>
<td>500</td>
<td>21</td>
<td>0.00</td>
<td>271</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>9</td>
<td>500</td>
<td>21</td>
<td>0.50</td>
<td>271</td>
<td>1.50</td>
<td>0.50</td>
</tr>
<tr>
<td>10</td>
<td>500</td>
<td>21</td>
<td>0.95</td>
<td>271</td>
<td>1.94</td>
<td>0.05</td>
</tr>
<tr>
<td>11</td>
<td>500</td>
<td>21</td>
<td>-0.95</td>
<td>3000</td>
<td>4.4e-4</td>
<td>1.7e-2</td>
</tr>
<tr>
<td>12</td>
<td>500</td>
<td>21</td>
<td>-0.50</td>
<td>3000</td>
<td>4.4e-3</td>
<td>1.3e-2</td>
</tr>
<tr>
<td>13</td>
<td>500</td>
<td>21</td>
<td>0.00</td>
<td>3000</td>
<td>8.9e-3</td>
<td>8.9e-3</td>
</tr>
<tr>
<td>14</td>
<td>500</td>
<td>21</td>
<td>0.50</td>
<td>3000</td>
<td>1.3e-2</td>
<td>4.4e-3</td>
</tr>
<tr>
<td>15</td>
<td>500</td>
<td>21</td>
<td>0.95</td>
<td>3000</td>
<td>1.7e-2</td>
<td>4.4e-4</td>
</tr>
</tbody>
</table>

Settings 16-20 use $r = 100,000$ and are effectively identical to $r = 3,000$. 
Comments on Experiment G

For large enough $r$, the true $\rho$ doesn't matter.

Regarding the oddity we noted earlier:

- When $r$ is large, we'd expect $\hat{\rho} = -1$ and $\hat{\rho} = +1$ about equally often, but $\hat{\rho} = +1$ is more frequent.
- In settings 11 through 20, $\hat{\rho} = +1$ occurs $\sim 1.5$ times as often as $\hat{\rho} = -1$, even when the true $\rho$ is $-0.95$.
- BUT this is not the case in setting 5 of Experiments C, E, and F, when $\rho = 0$ but $r$ is not so large and the predictor $\in (1, 10)$.

Questions: Is this an artifact of the model specification or of lmer? Does this matter, practically?
Conclusions about random regressions

Nothing the Experiments contradicted these first 4 hypotheses:

1. Predictor $\downarrow$, $\hat{\rho}$ is more likely to be $-1$; predictor $\uparrow$, $\hat{\rho}$ is less likely to be $-1$.
2. Given $N$, $s$, and $\rho$, as $r$ increases $\hat{\rho}$ is more likely to be $-1$.
3. Given $\rho$ and $r$, as either $N$ or $s$ increases, $\hat{\rho}$ is less likely to be $-1$.
4. Given $N$, $s$, and $r$, as $\rho$ goes to $-1$, $\hat{\rho}$ is more likely to be $-1$.

Experiments C, D, E, and F are consistent with Hypotheses 5 and 6:

5. Increased (decreased) $r$ is offset by increasing (decreasing) $N$, $s$.
6. Changing $r$ by about 2.5 is countered by changing $N$ by about 5 and $s$ by about 3.
7. Changes in $\rho$ affect the predictor less than changes in $N$ or $s$. If $r$ is big enough, for any $\rho$, $\hat{\rho}$ is very likely to be $\pm 1$.

Re “Changes in $\rho$ affect the predictor less than changes in $N$ or $s” — the experiments say “you asked the question poorly”.

For $N, s, r \ni$ some datasets give “bad” estimates, you need large changes in $\rho$ to get effects produced by moderate changes in $N, s, r$.

But large changes in $\rho$ also change the type of “bad” estimate; in that sense the predictor is misleading about $\hat{\rho}$’s behavior.

Another way to say this:

- $N, s$, and $r$ have simple, intuitive effects on the design’s resolution.
- $\rho$ has more complicated effects that depend on $N, s$, and $r$. 
We produced useful facts and intuition using with math & computing exercises that a capable MS student could do with faculty supervision.

Think how much useful information we could produce if we directed our effort toward such exercises and away from garbage like consistency theorems.