See chapter 1 in Regression and Other Stories.

Widen the cells.

```
html"""
<style>
    main {
        margin: 0 auto;
        max-width: 2000px;
        padding-left: max(160px, 10%);
        padding-right: max(160px, 10%);
    }
</style>
"""
```

A typical set of Julia packages to include in notebooks.

```
using Pkg ✓ , DrWatson ✓
```

```
begin
    # Specific to this notebook
    using GLM ✓

    # Specific to ROSStanPluto
    using StanSample ✓

    # Graphics related
    using CairoMakie ✓
    using AlgebraOfGraphics ✓

    # Include basic packages
    using RegressionAndOtherStories ✓
end
```

1.1 The three challenges of statistics.

Note

It is not common for me to copy from the book but this particular section deserves an exception!

The three challenges of statistical inference are:

1. Generalizing from sample to population, a problem that is associated with survey sampling but actually arises in nearly every application of statistical inference;
2. Generalizing from treatment to control group, a problem that is associated with causal inference, which is implicitly or explicitly part of the interpretation of most regressions we have seen; and
3. Generalizing from observed measurements to the underlying constructs of interest, as most of the time our data do not record exactly what we would ideally like to study.

All three of these challenges can be framed as problems of prediction (for new people or new items that are not in the sample, future outcomes under different potentially assigned treatments, and underlying constructs of interest, if they could be measured exactly).

1.2 Why learn regression?

hibbs =

	year	growth	vote	inc_party_candidat
1	1952	2.4	44.6	"Stevenson"
2	1956	2.89	57.76	"Eisenhower"
3	1960	0.85	49.91	"Nixon"
4	1964	4.21	61.34	"Johnson"
5	1968	3.02	49.6	"Humphrey"
6	1972	3.62	61.79	"Nixon"
7	1976	1.08	48.95	"Ford"
8	1980	-0.39	44.7	"Carter"
9	1984	3.86	59.17	"Reagan"
10	1988	2.27	53.94	"Bush, Sr."
⋮ more				
16	2012	0.95	52.0	"Obama"

```
hibbs =
CSV.read(ros_datadir("ElectionsEconomy",
"hibbs.csv"), DataFrame)
```

```
hibbs_lm =
StatsModels.TableRegressionModel{LinearModel{GLM

vote ~ 1 + growth

Coefficients:
────────────────────────────────────────────────────
                 Coef.  Std. Error      t  Pr(>|t
────────────────────────────────────────────────────
(Intercept)  46.2476      1.62193   28.51    <1e-1
growth        3.06053     0.696274   4.40    0.000
────────────────────────────────────────────────────
```

```
hibbs_lm = lm(@formula(vote ~ growth),
hibbs)
```

```
[-8.99292, 2.66743, 1.0609, 2.20753, -5.89044, 4
```

```
residuals(hibbs_lm)
```

```
2.2744434224582912
```

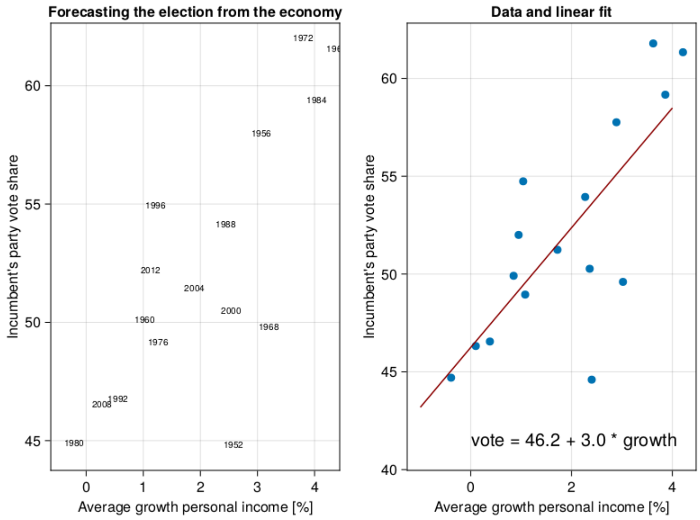
```
mad(residuals(hibbs_lm))
```

3.635681268522063

```
std(residuals(hibbs_lm))
```

[46.2476, 3.06053]

```
coef(hibbs_lm)
```




```
let
    fig = Figure()
    hibbs.label = string.(hibbs.year)
    xlabel = "Average growth personal
    income [%]"
    ylabel = "Incumbent's party vote share"
    let
        title = "Forecasting the election
        from the economy"
        ax = Axis(fig[1, 1]; title, xlabel,
        ylabel)
        for (ind, yr) in
        enumerate(hibbs.year)
            annotations!("$(yr)"; position=
            (hibbs.growth[ind],
            hibbs.vote[ind]), textsize=10)
        end
    end
    let
        x = LinRange(-1, 4, 100)
        title = "Data and linear fit"
        ax = Axis(fig[1, 2]; title, xlabel,
        ylabel)
        scatter!(hibbs.growth, hibbs.vote)
        lines!(x, coef(hibbs_lm)[1] .+
        coef(hibbs_lm)[2] .* x;
        color=:darkred)
        annotations!("vote = 46.2 + 3.0 *
        growth"; position=(0, 41))
    end
    fig
end
```

1.3 Some examples of regression.

Electric company

	post_test	pre_test	grade	t
1	48.9	13.8	1	1
2	70.5	16.5	1	1
3	89.7	18.5	1	1
4	44.2	8.8	1	1
5	77.5	15.3	1	1
6	84.7	15.0	1	1
7	78.9	19.4	1	1
8	86.8	15.0	1	1
9	60.8	11.8	1	1
10	75.7	16.4	1	1
⋮ more				
192	110.0	102.6	4	0

```
begin
    electric =
    CSV.read(ros_datadir("ElectricCompany",
    "electric.csv"), DataFrame)
    electric = electric[:, [:post_test,
    :pre_test, :grade, :treatment]]
    electric.grade =
    categorical(electric.grade)
    electric.treatment =
    categorical(electric.treatment)
    electric
end
```

A quick look at the overall values of `pre_test` and `post_test`.

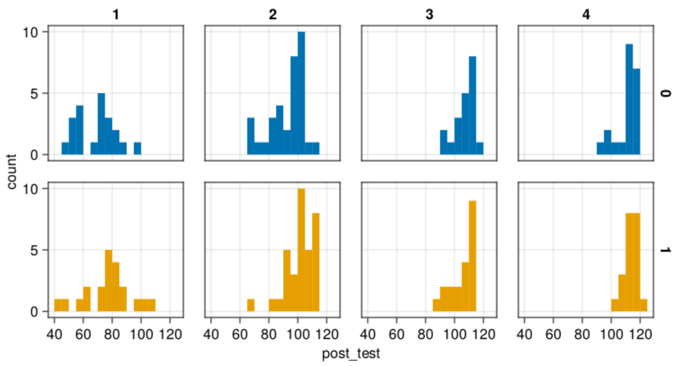
	variable	mean	min	median	max
1	:post_test	97.1495	44.2	102.3	122.0
2	:pre_test	72.2245	8.8	80.75	119.8
3	:grade	nothing	1	nothing	4
4	:treatment	nothing	0	nothing	1

```
describe(electric)
```

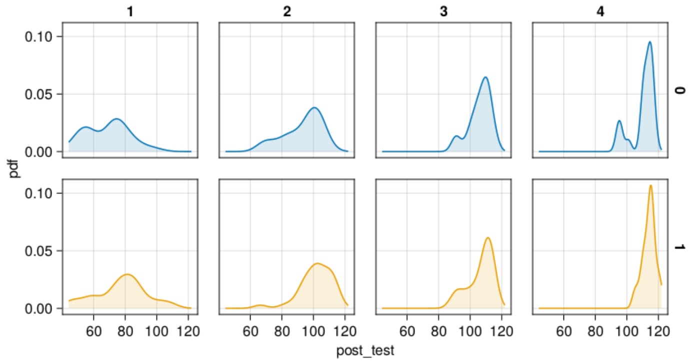
```
true
```

```
all(completecases(electric)) == true
```

Post-test density for each grade conditioned on treatment.



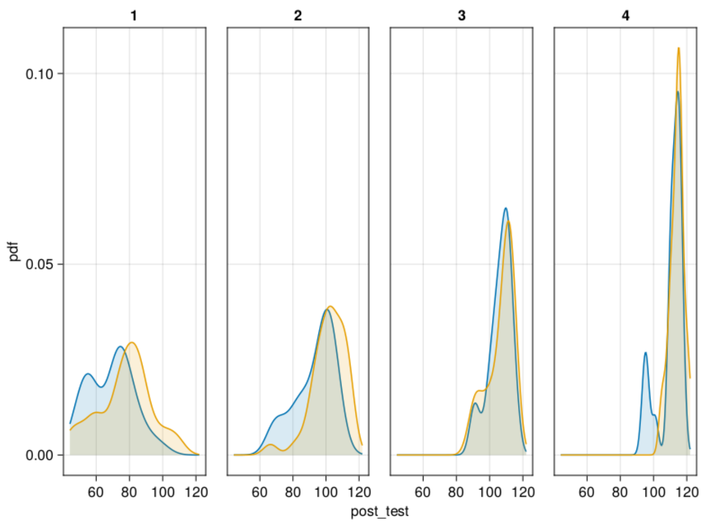

```
let
    f = Figure()
    axis = (; width = 150, height = 150)
    el = data(electric) *
    mapping(:post_test, col=:grade,
    color=:treatment)
    plt = el *
    AlgebraOfGraphics.histogram(;bins=20) *
    mapping(row=:treatment)
    draw!(f[1, 1], plt; axis)
    f
end
```



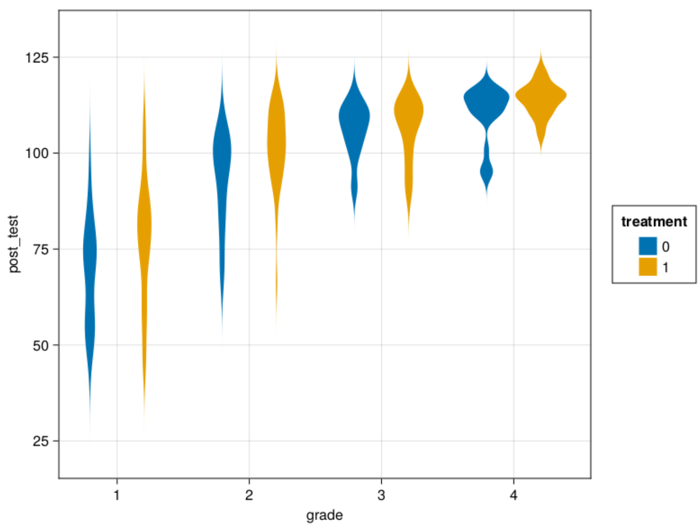

```
let
    f = Figure()
    axis = (; width = 150, height = 150)
    el = data(electric) *
    mapping(:post_test, col=:grade,
    color=:treatment)
    plt = el * AlgebraOfGraphics.density()
    * mapping(row=:treatment)
    draw!(f[1, 1], plt; axis)
    f
end
```

Note

In above cell, as density() is exported by both GLMakie and AlgebraOfGraphics, it needs to be qualified.




```
let
    f = Figure()
    el = data(electric) *
    mapping(:post_test, col=:grade)
    plt = el * AlgebraOfGraphics.density()
    * mapping(color=:treatment)
    draw!(f[1, 1], plt)
    f
end
```




```
let
    plt = data(electric) * visual(Violin) *
    mapping(:grade, :post_test,
    dodge=:treatment, color=:treatment)
    draw(plt)
end
```

Peacekeeping

peace =

	war	cfdate	faildate
1	"Afghanistan-Mujahideen"	8150	8257
2	"Afghanistan-Taliban"	8466	8505
3	"Algeria-FIS/AIS"	10149	12783
4	"Angola"	7820	8319
5	"Angola"	9089	10564
6	"Azerbaijan-N.K."	8643	8678
7	"Azerbaijan-N.K."	8901	12783
8	"Bangladesh-CHT"	8248	12783
9	"Myanmar-Karen"	8153	9282
10	"Myanmar-Karen"	9296	9907
⋮ more			
96	"Yugoslavia-Kosovo"	10751	12783

```
peace =
CSV.read(ros_datadir("PeaceKeeping",
"peacekeeping.csv"), missingstring="NA",
DataFrame)
```

	variable	mean	min
1	:war	nothing	"Afghanistan-Mujah:
2	:cfdate	8925.1	6985
3	:faildate	10795.8	7074
4	:peacekeepers	0.354167	0
5	:badness	-8.15228	-12.26
6	:delay	5.12177	0.04
7	:censored	0.416667	0

```
describe(peace)
```

A quick look at this Dates stuff!

```
8150
```
```julia
peace.cfdate[1]
```

```
1992-04-25T00:00:00
```
```julia
DateTime(1992, 4, 25)
```

```
107 days
```
```julia
Date(1992, 8, 10) - Date(1992, 4, 25)
```

```
1970-01-01
```
```julia
Date(1970,1,1)
```

```
1992-04-25
```
```julia
Date(1970,1,1) + Dates.Day(8150)
```

```
8150 days
```
```julia
Date(1992, 4, 25) - Date(1970, 1, 1)
```

```
107
```
```julia
peace.faildate[1] - peace.cfdate[1]
```

```julia
begin
    pks_df = peace[peace.peacekeepers .==
    1, [:cfdate, :faildate]]
    nopks_df = peace[peace.peacekeepers .==
    0, [:cfdate, :faildate]]
end;
```

```
0.4166666666666667
```
```julia
mean(peace.censored)
```

```
64
```
```julia
length(unique(peace.war))
```

```
0.5588235294117647
```
```julia
mean(peace[peace.peacekeepers .== 1,
:censored])
```

```
0.3387096774193548
```
```julia
mean(peace[peace.peacekeepers .== 0,
:censored])
```

```
1.382
```

```julia
mean(peace[peace.peacekeepers .== 1 .&&
peace.censored .== 0, :delay])
```

```
1.5153658536585364
```

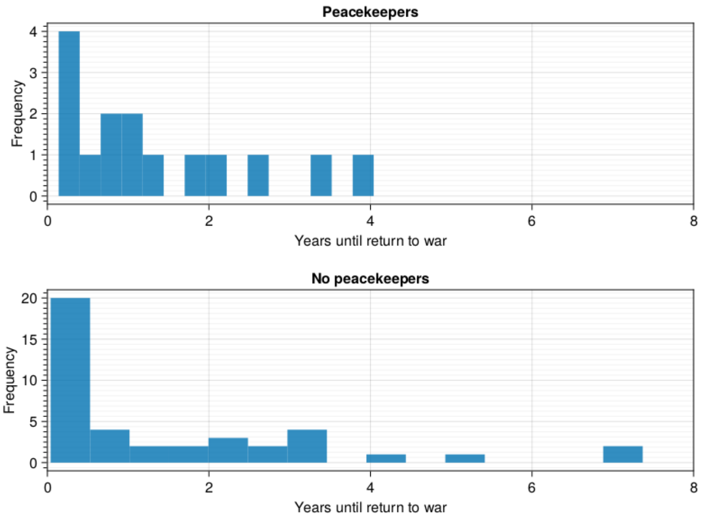
```julia
mean(peace[peace.peacekeepers .== 0 .&&
peace.censored .== 0, :delay])
```

```
1.05
```

```julia
median(peace[peace.peacekeepers .== 1 .&&
peace.censored .== 0, :delay])
```

```
0.59
```

```julia
median(peace[peace.peacekeepers .== 0 .&&
peace.censored .== 0, :delay])
```




```julia
let
    f = Figure()
    pks = peace[peace.peacekeepers .== 1
    .&& peace.censored .== 0, :]
    nopks = peace[peace.peacekeepers .== 0
    .&& peace.censored .== 0,:]

    for i in 1:2
        title = i == 1 ? "Peacekeepers" :
        "No peacekeepers"

        ax = Axis(f[i, 1]; title,
        xlabel="Years until return to war",
        ylabel = "Frequency",
    yminorticksvisible = true,
        yminorgridvisible = true,
        yminorticks = IntervalsBetween(8))

        xlims!(ax, [0, 8])
        hist!(i == 1 ? pks.delay :
        nopks.delay)
    end
    f
end
```

Note

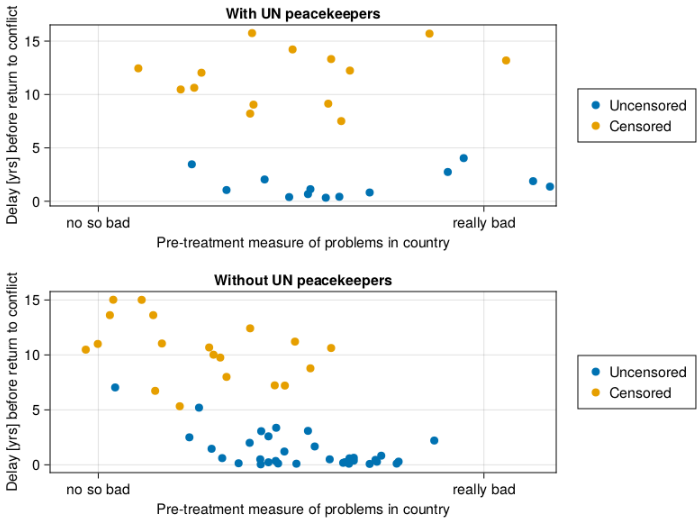
Censored means conflict had not returned until end of observation period (2004).

```julia
begin
    # Filter out missing badness rows.
    pb = peace[peace.badness .!== missing,
    :];

    # Delays until return to war for
    uncensored, peacekeeper cases
    pks_uc = pb[pb.peacekeepers .== 1 .&&
    pb.censored .== 0, :delay]
    # Delays until return to war for
    censored, peacekeeper cases
    pks_c = pb[pb.peacekeepers .== 1 .&&
    pb.censored .== 1, :delay]

    # No peacekeepr cases.
    nopks_uc = pb[pb.peacekeepers .== 0 .&&
    pb.censored .== 0, :delay]
    nopks_c = pb[pb.peacekeepers .== 0 .&&
    pb.censored .== 1, :delay]

    # Crude measure (:badness) used for
    assessing situation
    badness_pks_uc = pb[pb.peacekeepers .==
    1 .&& pb.censored .== 0,
        :badness]
    badness_pks_c = pb[pb.peacekeepers .== 1
     .&& pb.censored .== 1,
        :badness]
    badness_nopks_uc = pb[pb.peacekeepers
    .== 0 .&& pb.censored .== 0,
        :badness]
    badness_nopks_c = pb[pb.peacekeepers
    .== 0 .&& pb.censored .== 1,
        :badness]
end;
```




```julia
begin
    f = Figure()
    ax = Axis(f[1, 1], title = "With UN
    peacekeepers",
        xlabel = "Pre-treatment measure of
        problems in country",
        ylabel = "Delay [yrs] before return
        to conflict")
    sca1 = scatter!(badness_pks_uc, pks_uc)
    sca2 = scatter!(badness_pks_c, pks_c)
    xlims!(ax, [-13, -2.5])
    Legend(f[1, 2], [sca1, sca2],
    ["Uncensored", "Censored"])
    ax.xticks = ([-12, -4], ["no so bad",
    "really bad"])


    ax = Axis(f[2, 1], title = "Without UN
    peacekeepers",
        xlabel = "Pre-treatment measure of
        problems in country",
        ylabel = "Delay [yrs] before return
        to conflict")
    sca1 = scatter!(badness_nopks_uc,
    nopks_uc)
    sca2 = scatter!(badness_nopks_c,
    nopks_c)
    xlims!(ax, [-13, -2.5])
    Legend(f[2, 2], [sca1, sca2],
    ["Uncensored", "Censored"])
    ax.xticks = ([-12, -4], ["no so bad",
    "really bad"])

    f
end
```

1.4 Challenges in building, understanding, and interpreting regression.

Simple causal

Note

In models like below I usually prefer to create 2 separate Stan Language models, one for the continuous case and another for the binary case. But they can be combined in a single model as shown below. I'm using this example to show one way to handle vectors returned from Stan's cmdstan.

```
stan1_4_1 = "
data {
    int N;
    vector[N] x;
    vector[N] x_binary;
    vector[N] y;
}
parameters {
    vector[2] a;
    vector[2] b;
    vector<lower=0>[2] sigma;
}
model {
    // Priors
    a ~ normal(10, 10);
    b ~ normal(10, 10);
    sigma ~ exponential(1);
    // Likelihood
    y ~ normal(a[1] + b[1] * x, sigma[1]);
    y ~ normal(a[2] + b[2] * x_binary,
    sigma[2]);
}
";
```

Note

Aki Vehtari did not include a seed number in his code.

```
begin
    Random.seed!(123)
    n = 50
    x = rand(Uniform(1, 5), n)
    x_binary = [x[i] < 3 ? 0 : 1 for i in
    1:n]
    y = [rand(Normal(10 + 3x[i], 3), 1)[1]
    for i in 1:n]
end;
```

	parameters	mean	mcse	std	5⁽
1	"a[1]"	9.4	0.029	1.4	7.1
2	"a[2]"	16.0	0.013	0.69	15.(
3	"b[1]"	3.2	0.009	0.44	2.5
4	"b[2]"	7.0	0.02	1.0	5.3
5	"sigma[1]"	3.5	0.0056	0.34	3.0
6	"sigma[2]"	3.7	0.0062	0.37	3.1

```
let
    data = (N = n, x = x, x_binary =
    x_binary, y = y)
    global m1_4_1s = SampleModel("m1_4_1s",
    stan1_4_1);
    global rc1_4_1s = stan_sample(m1_4_1s;
    data)
    success(rc1_4_1s) && describe(m1_4_1s)
end
```

```
/var/folders/l7/pr04h0650q5dvqttnvs8s2c00000gn/T
ted.
```

Note

This is a good point to take a quick look at Pluto cell metadata: the top left `eye` symbol and the top right `3-dots in a circle` glyph (both only visible when the curser is in the input cell). Both are used quite often in these notebooks. Try them out!

The output of above method of the function `model_summary(::SampleModel)`, called directly on a SampleModel, is different from method `model_summary(::DataFrame)`, typically used later on. Above table shows important mcmc diagnostic columns like `n_eff` and `r_hat`.

If Stan parameters are vectors (as in this example), cmdstan returns those using '.' notation, e.g. a.1, a.2, ...

	parameters	median	mad_sd	mean	st
1	"a.1"	9.402	1.396	9.387	1.41
2	"a.2"	16.121	0.674	16.116	0.69
3	"b.1"	3.241	0.444	3.249	0.43
4	"b.2"	7.001	1.027	7.006	1.03
5	"sigma.1"	3.423	0.324	3.452	0.34
6	"sigma.2"	3.656	0.35	3.69	0.36

```
if success(rc1_4_1s)
    post1_4_1s = read_samples(m1_4_1s,
    :dataframe)
    model_summary(post1_4_1s,
    names(post1_4_1s))
end
```

With vector parameters `read_samples()` can create a nested DataFrame:

nd1_4_1s =

	a	b
1	[8.24802, 16.424]	[3.68005, 7.67235
2	[8.83781, 16.203]	[3.34579, 5.17041
3	[9.84559, 16.0788]	[3.35379, 5.76383
4	[11.0378, 16.134]	[3.05522, 6.06792
5	[10.8772, 15.807]	[3.00888, 5.84699
6	[10.6881, 16.2392]	[2.72581, 6.47463
7	[7.44081, 16.0745]	[3.96654, 6.36506
8	[10.062, 15.7894]	[2.99653, 7.15061
9	[9.95532, 15.5407]	[3.10871, 6.93601
10	[11.3043, 16.2698]	[2.75944, 5.98389
⋮ more		
4000	[10.5984, 17.5858]	[2.70611, 4.93714

```
nd1_4_1s = read_samples(m1_4_1s,
:nesteddataframe)
```

ms1_4_1s =

	parameters	median	mad_sd	mean	st
1	"a.1"	9.402	1.396	9.387	1.4
2	"a.2"	16.121	0.674	16.116	0.6
3	"b.1"	3.241	0.444	3.249	0.4
4	"b.2"	7.001	1.027	7.006	1.0
5	"sigma.1"	3.423	0.324	3.452	0.3
6	"sigma.2"	3.656	0.35	3.69	0.3

```
ms1_4_1s = success(rc1_4_1s) &&
model_summary(post1_4_1s, names(post1_4_1s))
```

1.027

```
ms1_4_1s["b.2", "mad_sd"]
```

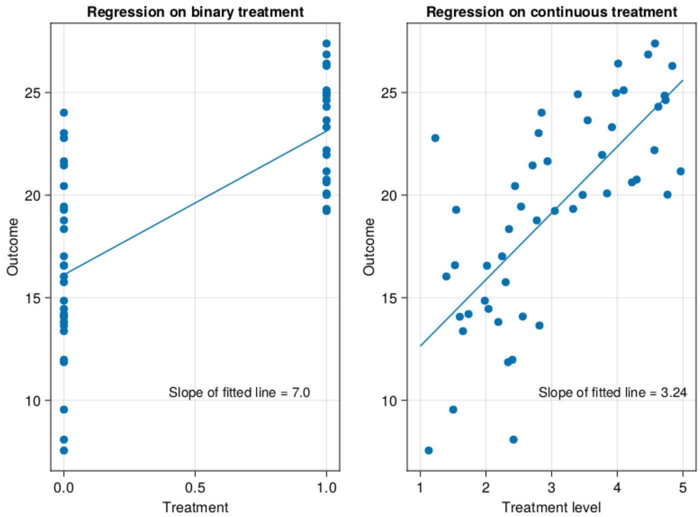
Nested dataframes are handy to obtain a matrix of say the b values:

```
4000×2 Matrix{Float64}:
 3.68005  7.67235
 3.34579  5.17041
 3.35379  5.76383
 3.05522  6.06792
 3.00888  5.84699
 2.72581  6.47463
 3.96654  6.36506
 ⋮
 3.24994  7.27068
 2.90944  5.6818
 3.25186  7.30268
 3.82635  7.82361
 3.5807   8.38818
 2.70611  4.93714
```

```
array(nd1_4_1s, :b)
```

```
4000×2 Matrix{Float64}:
 3.68005  7.67235
 3.34579  5.17041
 3.35379  5.76383
 3.05522  6.06792
 3.00888  5.84699
 2.72581  6.47463
 3.96654  6.36506
 ⋮
 3.24994  7.27068
 2.90944  5.6818
 3.25186  7.30268
 3.82635  7.82361
 3.5807   8.38818
 2.70611  4.93714
```

```
Array(post1_4_1s[:, ["b.1", "b.2"]])
```




```
let
    x1 = 1.0:0.01:5.0
    f = Figure()
    medians = ms1_4_1s[:, "median"]
    ax = Axis(f[1, 2], title = "Regression
    on continuous treatment",
        xlabel = "Treatment level", ylabel
        = "Outcome")
    sca1 = scatter!(x, y)
    annotations!("Slope of fitted line =
    $(round(medians[3], digits=2))",
        position = (2.8, 10), textsize=15)
    lin1 = lines!(x1, medians[1] .+
    medians[3] * x1)

    x2 = 0.0:0.01:1.0
    ax = Axis(f[1, 1], title="Regression on
    binary treatment",
        xlabel = "Treatment", ylabel =
        "Outcome")
    sca1 = scatter!(x_binary, y)
    lin1 = lines!(x2, medians[2] .+
    medians[4] * x2)
    annotations!("Slope of fitted line =
    $(round(medians[4], digits=2))",
        position = (0.4, 10), textsize=15)
    f
end
```

```
stan1_4_2 = "
data {
    int N;
    vector[N] x;
    vector[N] y;
}
parameters {
    vector[2] a;
    real b;
    real b_exp;
    vector<lower=0>[2] sigma;
}
model {
    // Priors
    a ~ normal(10, 5);
    b ~ normal(0, 5);
    b_exp ~ normal(5, 5);
    sigma ~ exponential(1);
    // Likelihood
    vector[N] mu;
    for ( i in 1:N )
        mu[i] = a[2] + b_exp * exp(-x[i]);
    y ~ normal(mu, sigma[2]);
    y ~ normal(a[1] + b * x, sigma[1]);
}
";
```

	parameters	mean	mcse	std	5
1	"a[1]"	13.0	0.02	0.98	11.(
2	"a[2]"	5.9	0.007	0.38	5.3
3	"b"	-1.74	0.01	0.3	-2.:
4	"b_exp"	17.94	0.06	3.02	12.:
5	"sigma[1]"	2.3	0.0038	0.23	1.9
6	"sigma[2]"	2.2	0.0038	0.23	1.8

```
let
    #Random.seed!(1533)
    n1 = 50
    x1 = rand(Uniform(1, 5), n1)
    y1 = [rand(Normal(5 + 30exp(-x1[i]),
    2), 1)[1] for i in 1:n]
    data = (N = n1, x = x1, y = y1)
    global m1_4_2s = SampleModel("m1.4_2s",
    stan1_4_2);
    global rc1_4_2s = stan_sample(m1_4_2s;
    data)
    success(rc1_4_2s) && describe(m1_4_2s)
end
```

```
/var/folders/l7/pr04h0650q5dvqttnvs8s2c00000gn/T
ted.
```

	parameters	median	mad_sd	mean	st
1	"a.1"	12.758	0.95	12.747	0.98
2	"a.2"	5.919	0.386	5.927	0.38
3	"b"	-1.743	0.285	-1.743	0.29
4	"b_exp"	17.931	2.967	17.936	3.02
5	"sigma.1"	2.266	0.225	2.281	0.23
6	"sigma.2"	2.17	0.224	2.186	0.22

```
if success(rc1_4_2s)
    post1_4_2s = read_samples(m1_4_2s,
    :dataframe)
    ms1_4_2s = model_summary(post1_4_2s,
    ["a.1", "a.2", "b", "b_exp", "sigma.1",
    "sigma.2"])
end
```

nd1_4_2s =

	b	b_exp	a
1	-2.00479	10.8794	▸[13.6309, 6.32045]
2	-1.48959	17.7817	▸[11.7728, 6.34427]
3	-1.42222	20.2764	▸[11.4523, 5.56977]
4	-2.08127	14.8388	▸[13.6406, 5.95356]
5	-1.75675	17.4339	▸[13.041, 6.42338]
6	-1.00498	18.4521	▸[10.4521, 6.09275]
7	-2.09687	18.1361	▸[13.6547, 5.40958]
8	-1.61418	16.5616	▸[13.3017, 6.63865]
9	-1.70031	16.8807	▸[13.174, 6.87947]
10	-2.02728	15.8727	▸[13.6482, 5.19274]
⋮ more			
4000	-1.97497	16.1358	▸[13.7143, 5.49498]

```
nd1_4_2s = read_samples(m1_4_2s,
:nesteddataframe)
```

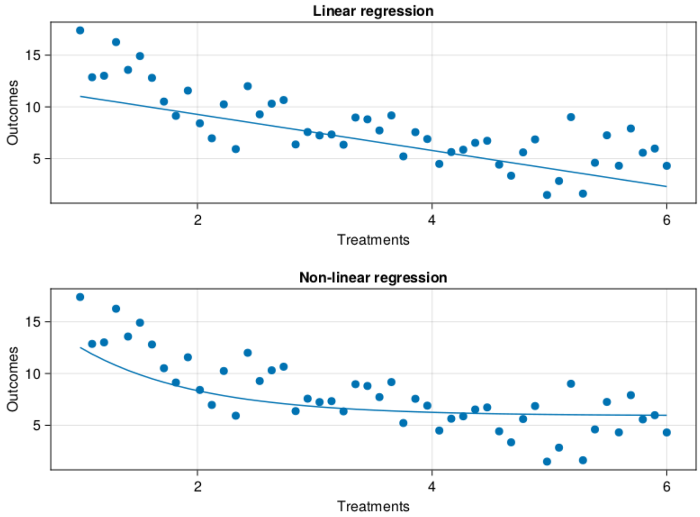
```
4000×2 Matrix{Float64}:
 13.6309   6.32045
 11.7728   6.34427
 11.4523   5.56977
 13.6406   5.95356
 13.041    6.42338
 10.4521   6.09275
 13.6547   5.40958
  ⋮
 11.4669   6.19831
 13.0419   5.50892
 13.1873   5.30649
 14.6559   6.19775
 10.8018   6.18113
 13.7143   5.49498
```

```
array(nd1_4_2s, :a)
```

```
â₁, â₂, b̂, b̂ₑₓₚ, σ̂₁, σ̂₂ = [ms1_4_2s[p,
"median"] for p in ["a.1", "a.2", "b",
"b_exp", "sigma.1", "sigma.2"]];
```

5.919

```
â₂
```




```
let
    x1 = LinRange(1, 6, 50)
    y1 = [rand(Normal(5 + 30exp(-x1[i]),
    2), 1)[1] for i in 1:length(x1)]

    f = Figure()
    ax = Axis(f[1, 1], title = "Linear
    regression",
        xlabel = "Treatments", ylabel =
        "Outcomes")
    scatter!(x1, y1)
    lines!(x1, â₁ .+ b̂ .* x1)

    ax = Axis(f[2, 1], title = "Non-linear
    regression",
        xlabel = "Treatments", ylabel =
        "Outcomes")
    scatter!(x1, y1)
    lines!(x1, â₂ .+ b̂ₑₓₚ .* exp.(-x1))
    f
end
```

	xx	z	yy
1	3.1425	0	37.5294
2	0.0335661	1	30.0628
3	1.60408	0	28.1095
4	0.478735	1	31.3638
5	2.65874	0	35.7382
6	1.02705	1	36.0009
7	1.28799	0	24.3213
8	0.052966	1	29.5242
9	0.543994	0	25.4181
10	0.0304007	1	25.1863
⋮ more			
100	0.00156342	1	28.8751

```
begin
    Random.seed!(12573)
    n2 = 100
    z = repeat([0, 1]; outer=50)
    df1_8 = DataFrame()
    df1_8.xx = [(z[i] == 0 ? rand(Normal(0,
    1.2), 1).^2 : rand(Normal(0, 0.8),
    1).^2)[1] for i in 1:n2]
    df1_8.z = z
    df1_8.yy = [rand(Normal(20 .+
    5df1_8.xx[i] .+ 10df1_8.z[i], 3), 1)[1]
    for i in 1:n2]
    df1_8
end
```

```
lm1_8 =
StatsModels.TableRegressionModel{LinearModel{GLM…

yy ~ 1 + xx + z

Coefficients:
─────────────────────────────────────────────────
                 Coef.   Std. Error      t  Pr(>|t
─────────────────────────────────────────────────
(Intercept)  20.1093      0.529823  37.95    <1e-5
xx            4.97503     0.213492  23.30    <1e-4
z             9.625       0.604978  15.91    <1e-2
─────────────────────────────────────────────────
```

```
lm1_8 = lm(@formula(yy ~ xx + z), df1_8)
```

```
lm1_8_0 =
StatsModels.TableRegressionModel{LinearModel{GLM…

yy ~ 1 + xx

Coefficients:
─────────────────────────────────────────────────
                 Coef.   Std. Error      t  Pr(>|t
─────────────────────────────────────────────────
(Intercept)  20.0337      0.544062  36.82    <1e-3
xx            5.01957     0.226965  22.12    <1e-2
─────────────────────────────────────────────────
```

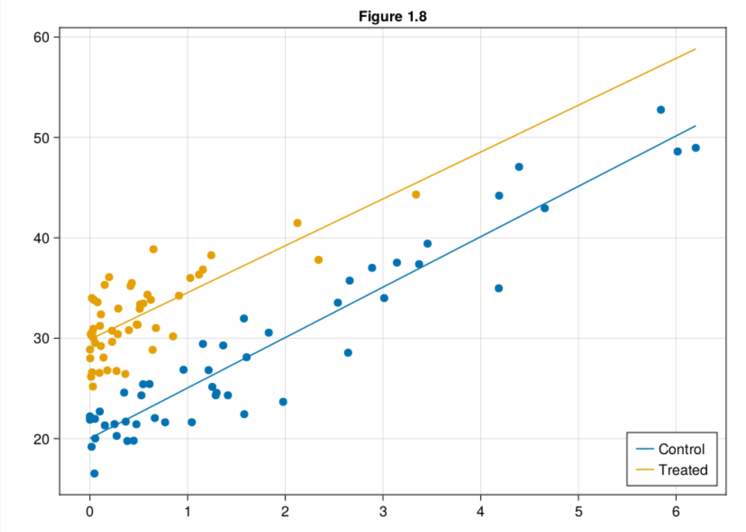
```
lm1_8_0 = lm(@formula(yy ~ xx),
df1_8[df1_8.z .== 0, :])
```

```
lm1_8_1 =
StatsModels.TableRegressionModel{LinearModel{GLM…

yy ~ 1 + xx

Coefficients:
─────────────────────────────────────────────────
                 Coef.   Std. Error      t  Pr(>|t
─────────────────────────────────────────────────
(Intercept)  29.8841      0.49051   60.92    <1e-4
xx            4.66553     0.609796   7.65    <1e-0
─────────────────────────────────────────────────
```

```
lm1_8_1 = lm(@formula(yy ~ xx),
df1_8[df1_8.z .== 1, :])
```




```
let
    â₁, b̂₁ = coef(lm1_8_0)
    â₂, b̂₂ = coef(lm1_8_1)
    x = LinRange(0, maximum(df1_8.xx), 40)

    f = Figure()
    ax = Axis(f[1, 1]; title="Figure 1.8")
    scatter!(df1_8.xx[df1_8.z .== 0],
    df1_8.yy[df1_8.z .== 0])
    scatter!(df1_8.xx[df1_8.z .== 1],
    df1_8.yy[df1_8.z .== 1])
    lines!(x, â₁ .+ b̂₁ * x, label =
    "Control")
    lines!(x, â₂ .+ b̂₂ * x, label =
    "Treated")
    axislegend(; position=(:right, :bottom))
    current_figure()
end
```

1.5 Classical and Bayesian inference.

1.6 Computing least-squares and Bayesian regression.

1.8 Exercises.

Helicopters

```
helicopters =
```

	Helicopter_ID	width_cm	length_cm	time_s
1	1	4.6	8.2	1.64
2	1	4.6	8.2	1.74
3	1	4.6	8.2	1.68
4	1	4.6	8.2	1.62
5	1	4.6	8.2	1.68
6	1	4.6	8.2	1.7
7	1	4.6	8.2	1.62
8	1	4.6	8.2	1.66
9	1	4.6	8.2	1.69
10	1	4.6	8.2	1.62
⋮ more				
20	2	4.6	8.2	1.61

```
helicopters =
CSV.read(ros_datadir("Helicopters",
"helicopters.csv"), DataFrame)
```

Simulate 40 helicopters.

	width_cm	length_cm	time_sec
1	3.34899	0.821947	0.505001
2	2.78591	6.26776	1.33218
3	6.20855	10.1125	1.52182
4	8.01419	9.74486	1.53831
5	2.82847	8.63838	1.13026
6	4.21747	15.7541	1.86944
7	5.37202	9.76108	1.61338
8	6.95056	12.7485	1.79877
9	5.29706	16.2845	2.15528
10	1.71303	9.24657	1.36999
⋮ more			
40	6.74333	10.5377	1.76125

```
begin
    helis = DataFrame(width_cm =
    rand(Normal(5, 2), 40), length_cm =
    rand(Normal(10, 4), 40))
    helis.time_sec = 0.5 .+ 0.04 .*
    helis.width_cm .+ 0.08 .*
    helis.length_cm .+ 0.1 .*
    rand(Normal(0, 1), 40)
    helis
end
```

```
stan1_5 = "
data {
    int N;
    vector[N] w;
    vector[N] l;
    vector[N] y;
}
parameters {
    real a;
    real b;
    real c;
    real<lower=0> sigma;
}
model {
    // Priors
    a ~ normal(10, 5);
    b ~ normal(0, 5);
    sigma ~ exponential(1);

    // Likelihood time on width
    vector[N] mu;
    for ( i in 1:N )
        mu[i] = a + b * w[i] + c * l[i];
    y ~ normal(mu, sigma);
}
";
```

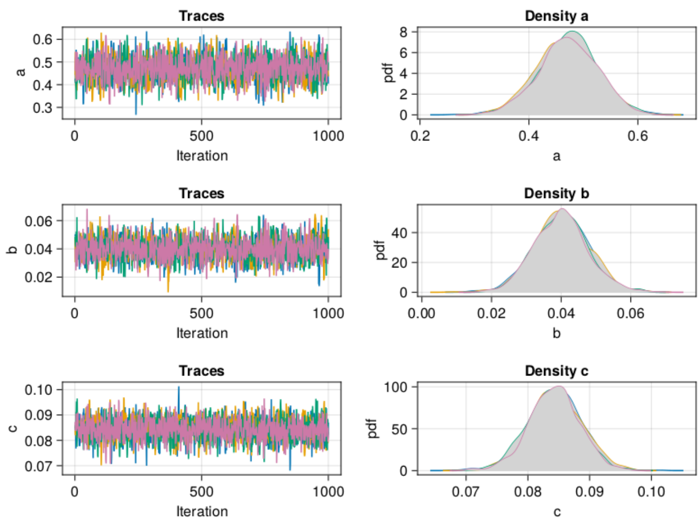
	parameters	mean	mcse	std
1	"a"	0.471211	0.00116792	0.05343
2	"b"	0.0400364	0.000170083	0.00764
3	"c"	0.0843282	8.34419e-5	0.00418
4	"sigma"	0.103305	0.000284722	0.01245

```
let
    data = (N = nrow(helis), y =
    helis.time_sec, w = helis.width_cm, l =
    helis.length_cm)
    global m1_5s = SampleModel("m1.5s",
    stan1_5);
    global rc1_5s = stan_sample(m1_5s; data)
    success(rc1_5s) && describe(m1_5s)
end
```

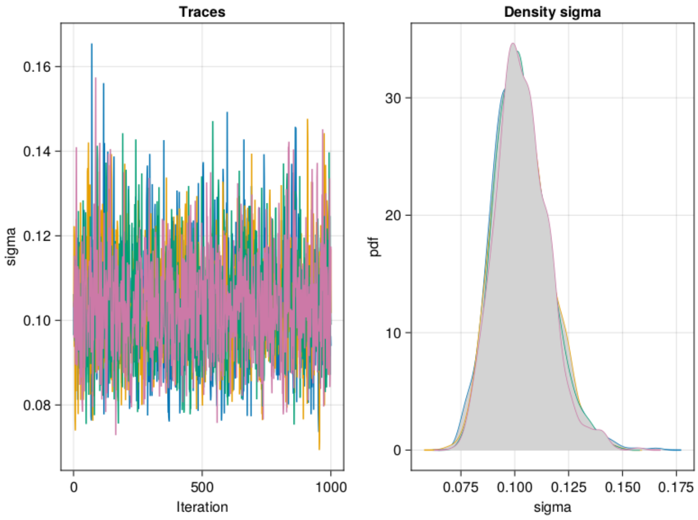
```
Informational Message: The current Metropolis
jected because of the following issue:
Exception: normal_lpdf: Scale parameter is 0,
r/folders/l7/pr04h0650q5dvqttnvs8s2c00000gn/T/
3, column 1 to column 23)
If this warning occurs sporadically, such as f
types like covariance matrices, then the sampl
but if this warning occurs often then your mod
conditioned or misspecified.
```

	parameters	median	mad_sd	mean	st
1	"a"	0.4724	0.0532	0.4712	0.05
2	"b"	0.04	0.0074	0.04	0.00
3	"c"	0.0843	0.0041	0.0843	0.00
4	"sigma"	0.1022	0.0123	0.1033	0.01

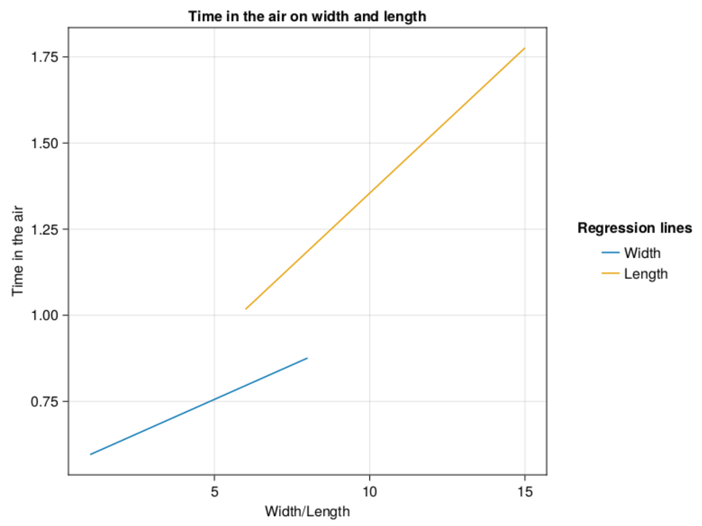
```
if success(rc1_5s)
    post1_5s = read_samples(m1_5s,
    :dataframe)
    model_summary(post1_5s, [:a, :b, :c,
    :sigma]; digits=4)
end
```




```
plot_chains(post1_5s, [:a, :b, :c])
```




```
plot_chains(post1_5s, [:sigma])
```




```
let
    w_range = LinRange(1.0, 8.0, 100)
    w_times = mean.(link(post1_5s, (r, w) -
    > r.a + r.c + r.b * w, w_range))
    l_range = LinRange(6.0, 15.0, 100)
    l_times = mean.(link(post1_5s, (r, l) -
    > r.a + r.b + r.c * l, l_range))

    f = Figure()
    ax = Axis(f[1, 1], title = "Time in the
    air on width and length",
        xlabel = "Width/Length", ylabel =
        "Time in the air")

    lines!(w_range, w_times; label="Width")
    lines!(l_range, l_times; label="Length")

    f[1, 2] = Legend(f, ax, "Regression
    lines", framevisible = false)

    current_figure()
end
```

lnk1_5s =

[[0.894687, 0.881022, 0.968542, 0.959784, 0.9846

```
lnk1_5s = link(post1_5s, (r, l) -> r.a + r.b
  + r.c * l, [5, 10,12])
```

[0.933737, 1.35465, 1.52305]

```
median.(lnk1_5s)
```

[0.038196, 0.03567, 0.0378395]

```
mad.(lnk1_5s)
```

[0.932888, 1.35453, 1.52319]

```
mean.(link(post1_5s, (r, l) -> r.a + r.b +
r.c * l, [5, 10,12]))
```

	a	b	c	sigma
1	0.424878	0.0489688	0.084168	0.10622
2	0.406936	0.0504471	0.0847278	0.096517
3	0.516133	0.0306989	0.084342	0.10076
4	0.531751	0.0441585	0.0767748	0.10188
5	0.554138	0.0364994	0.0788079	0.10753
6	0.535185	0.0398682	0.0791222	0.10382
7	0.495928	0.0451504	0.0801624	0.10314
8	0.458166	0.0361234	0.0874675	0.10522
9	0.450037	0.0333757	0.0870754	0.10653
10	0.451163	0.0434026	0.0841383	0.09661
⋮ more				
4000	0.506864	0.0397577	0.0804016	0.11744

```
read_samples(m1_5s, :nesteddataframe)
```

No nested columns found.