DATA SCIENCE TOOLS FOR PETROLEUM EXPLORATION AND PRODUCTION

Matteo Niccoli and Thomas Speidel

AGENDA

PART I (Python) – Statistically significant correlation

PART II (R)– Variable selection and multivariate analysis

AGENDA

PART I (Python) – Statistically significant correlation (linear, bivariate)

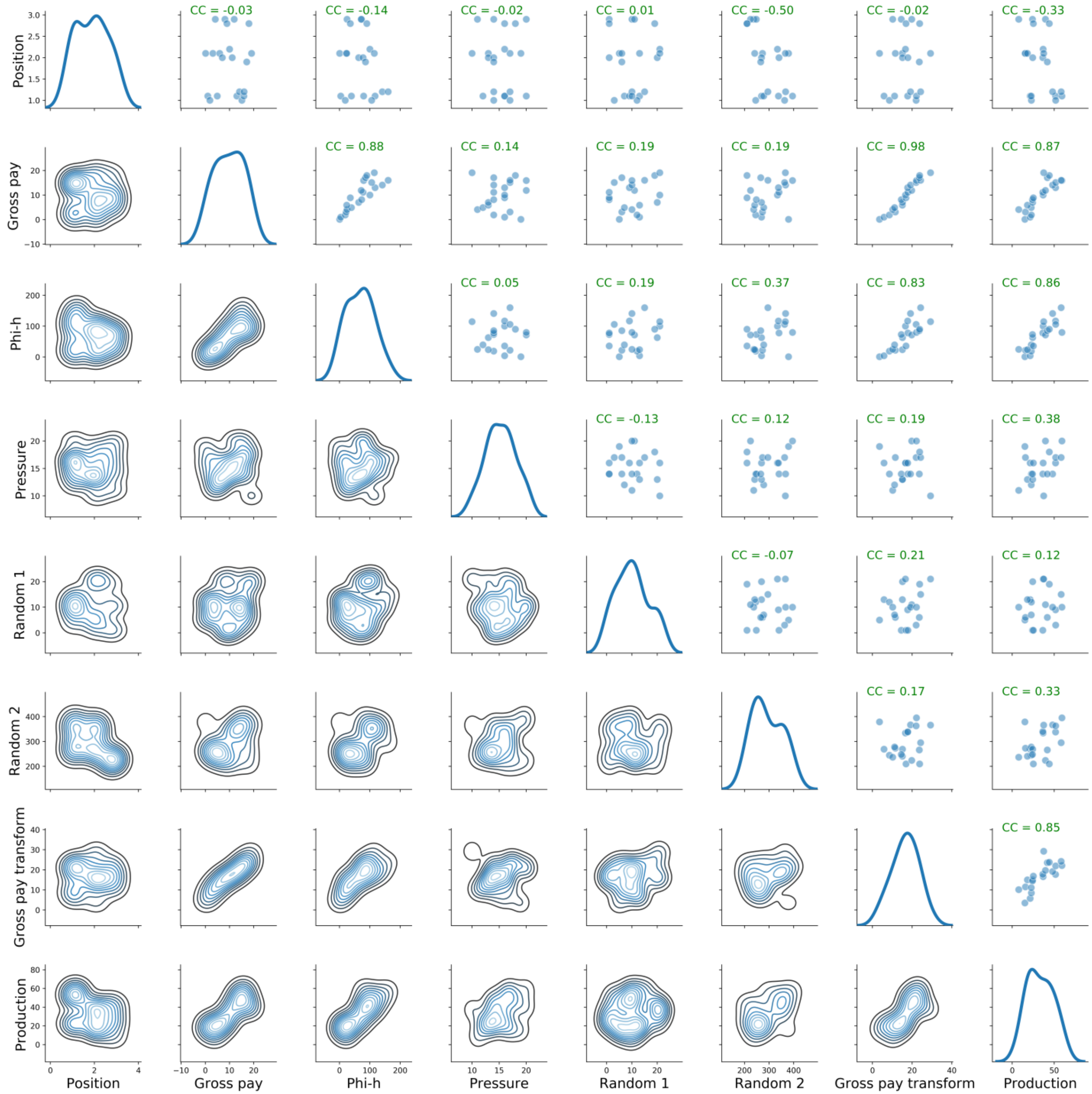
PART II (R)– Variable selection and multivariate analysis

When is a correlation statistically significant?

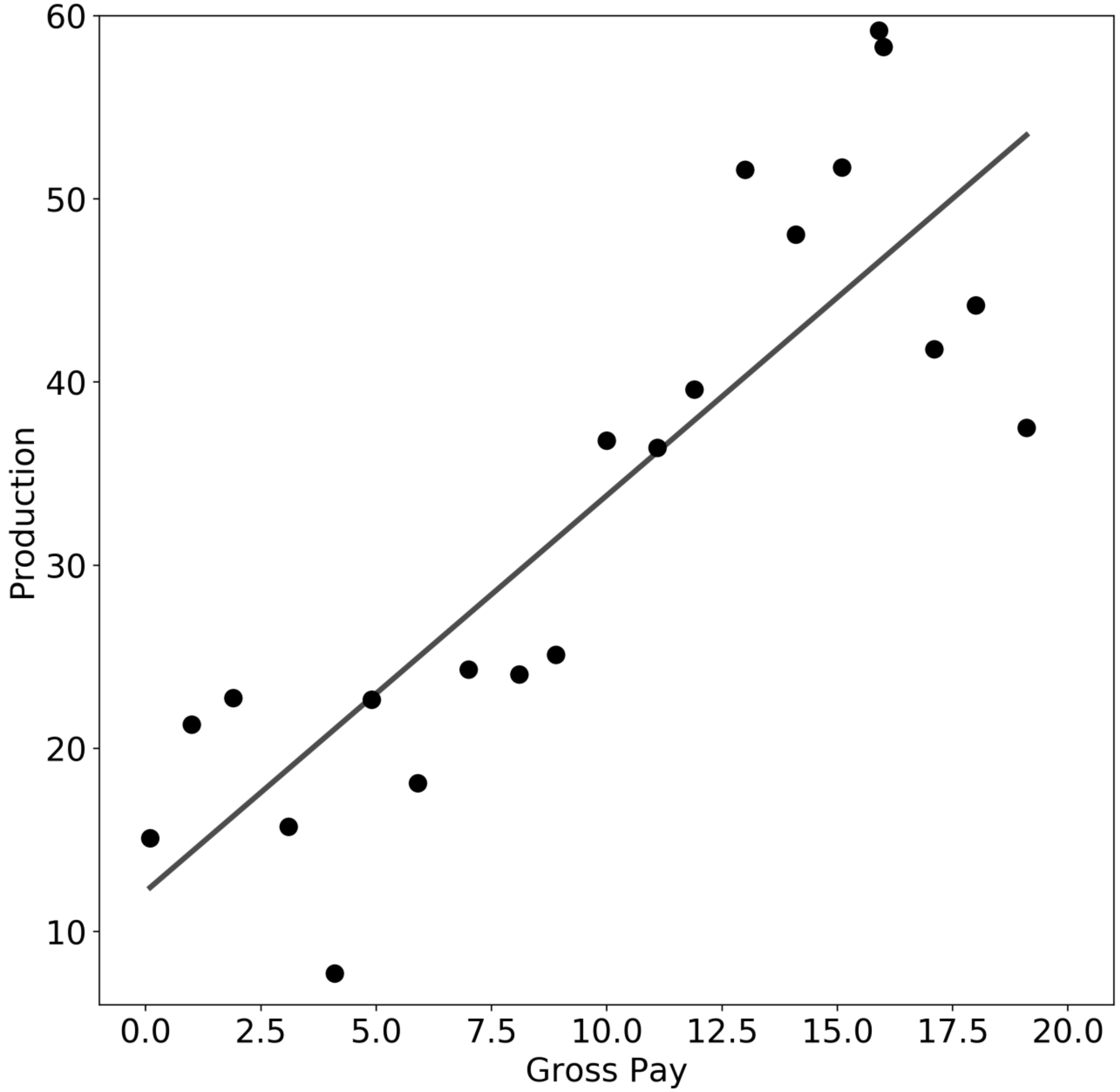
Well number	X1 Gross pay (m)	X2 Phi-h	X3 Position	X4 Pressure (MPa)	X5 - random R1	X6 - random R2	X7 - artificial Gross pay transform	Y Production (bbls/d*10)
1	0.1	0.5	2.1	19.0	5.0	379.0	3.5	15.1
2	1.0	4.0	1.1	16.0	13.0	269.0	5.8	21.3
3	1.9	19.0	1.0	14.0	12.0	245.0	8.5	22.8
4	3.1	21.7	2.1	17.0	6.0	273.0	11.5	15.7
5	4.1	24.6	2.9	11.0	10.0	237.0	10.2	7.7
6	4.9	39.2	1.1	12.0	7.0	278.0	11.1	22.7
7	5.9	23.6	2.1	13.0	13.0	241.0	15.0	18.1
8	7.0	63.0	2.0	13.0	20.0	269.0	15.1	24.3
9	8.1	72.9	2.9	14.0	1.0	248.0	14.5	24.0
10	8.9	35.6	2.8	16.0	1.0	210.0	16.9	25.1
11	10.0	100.0	2.2	16.0	21.0	334.0	16.6	36.8
12	11.1	77.7	2.0	14.0	1.0	340.0	17.8	36.4
13	11.9	71.4	2.9	20.0	11.0	224.0	19.7	39.6
14	13.0	117.0	1.1	16.0	9.0	338.0	17.7	51.6
15	14.1	141.0	1.2	14.0	10.0	367.0	19.2	48.1
16	15.1	105.7	1.0	17.0	3.0	363.0	22.0	51.7
17	15.9	79.5	1.1	20.0	10.0	395.0	22.2	59.2
18	16.0	160.0	1.2	17.0	15.0	295.0	24.2	58.3
19	17.1	85.5	1.9	14.0	6.0	266.0	23.6	41.8
20	18.0	90.0	2.8	18.0	19.0	210.0	23.8	44.2
21	19.1	114.6	2.1	10.0	21.0	366.0	29.3	37.5

Data from Hunt (2013)

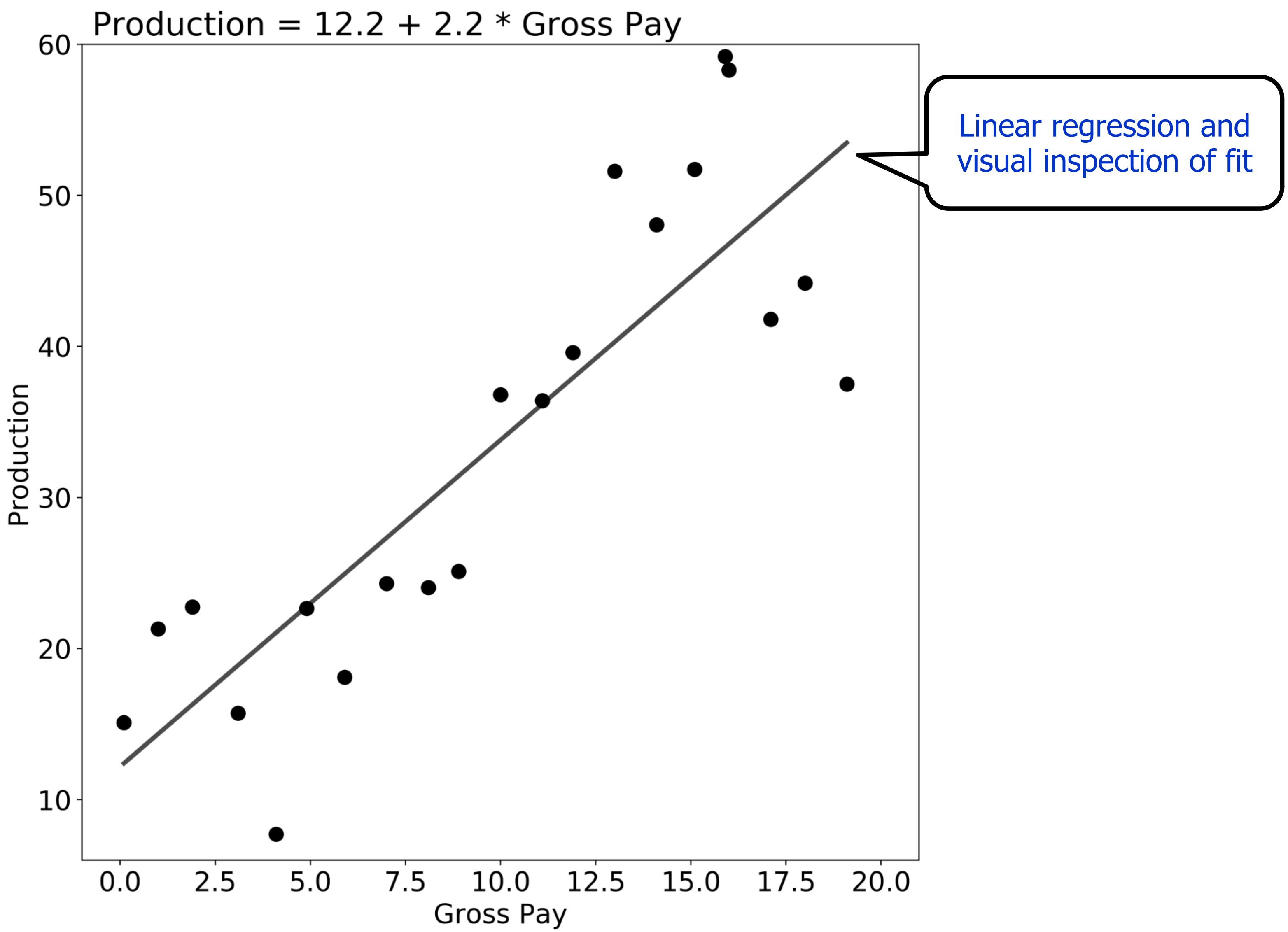
When is a correlation statistically significant?



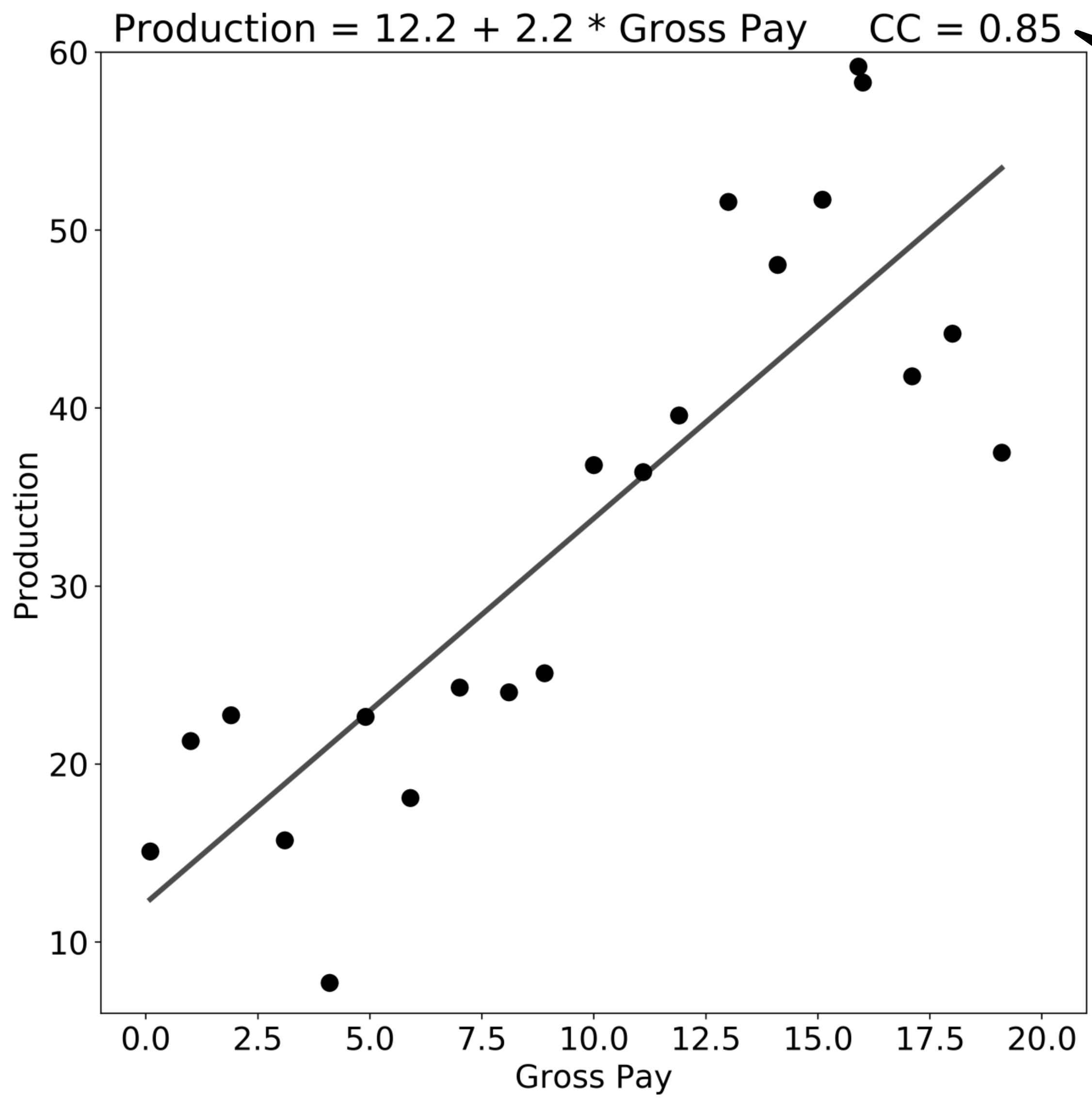
When is a correlation statistically significant?



When is a correlation statistically significant?

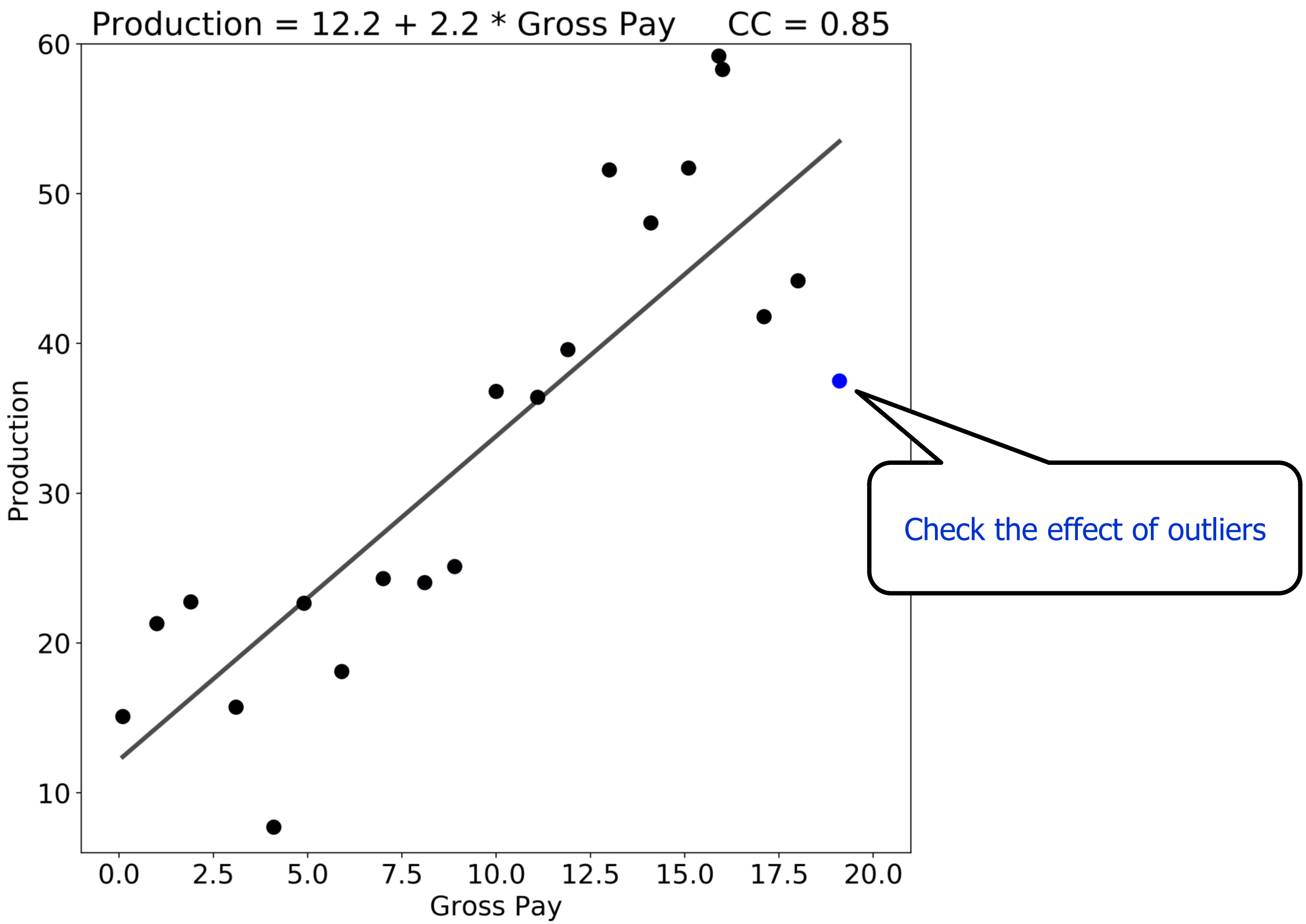


When is a correlation statistically significant?

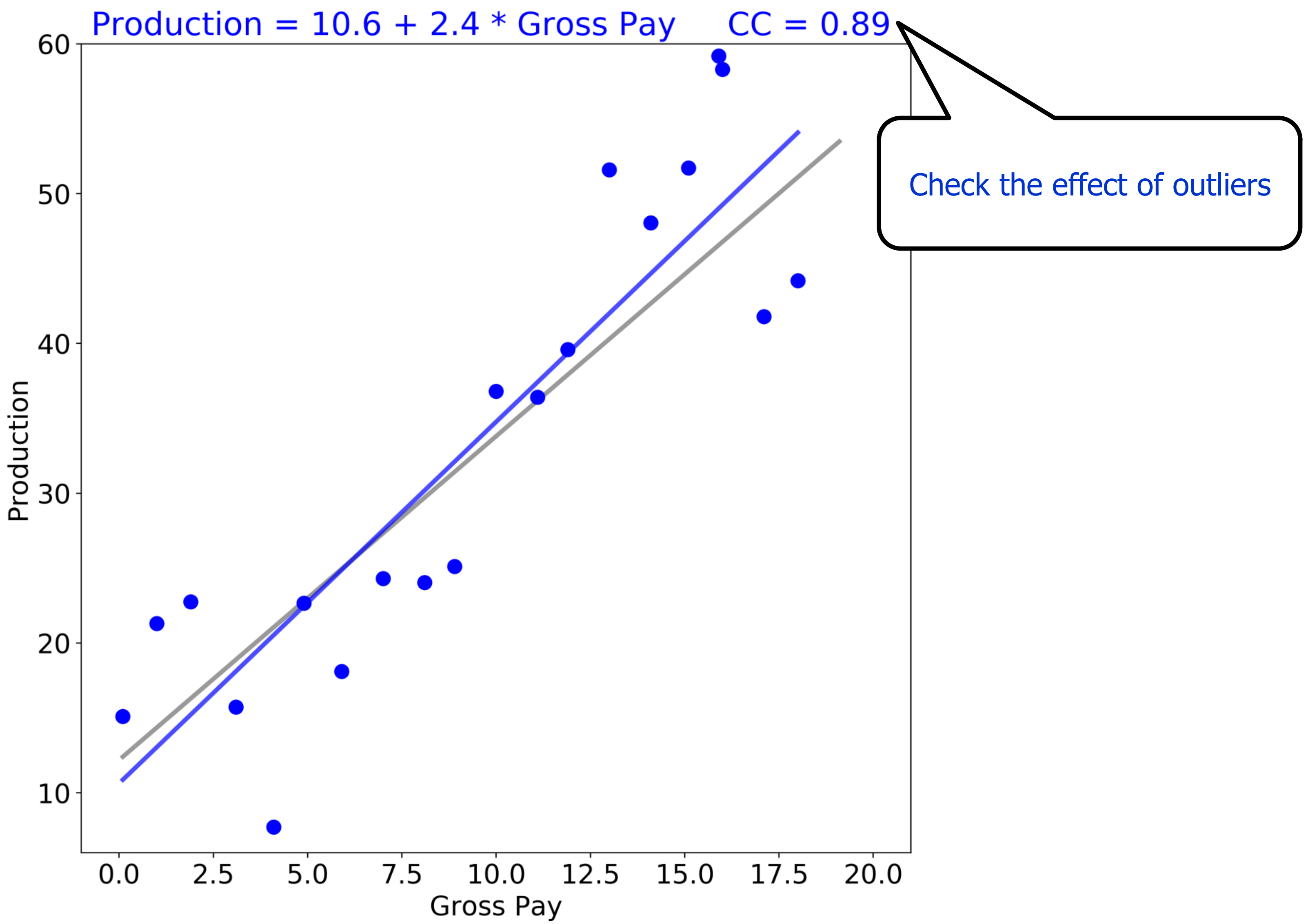


Correlation coefficient and / or goodness of fit

When is a correlation statistically significant?



When is a correlation statistically significant?



A deeper look

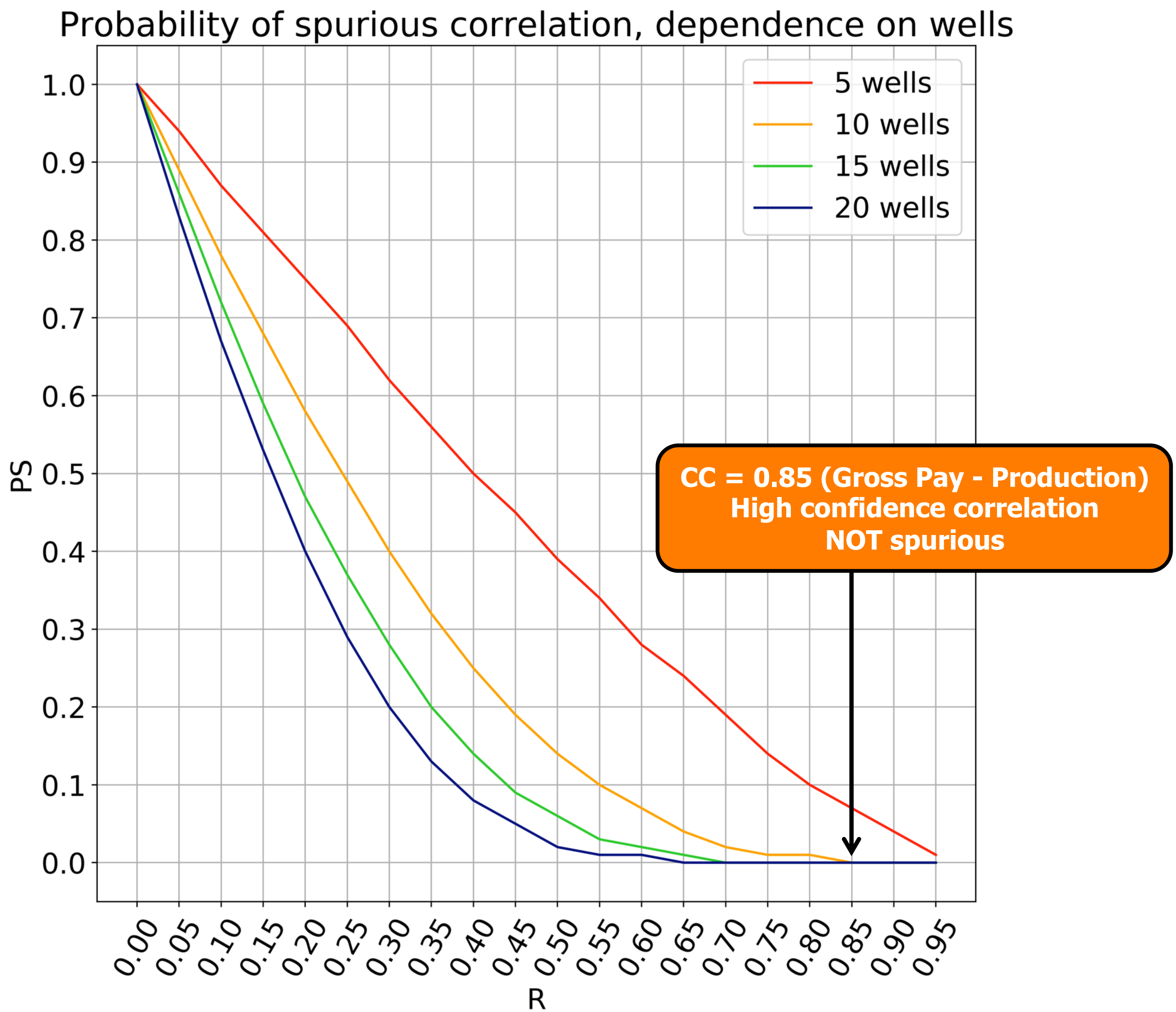
Probability of spurious correlation
(Kalkomey, 1997)

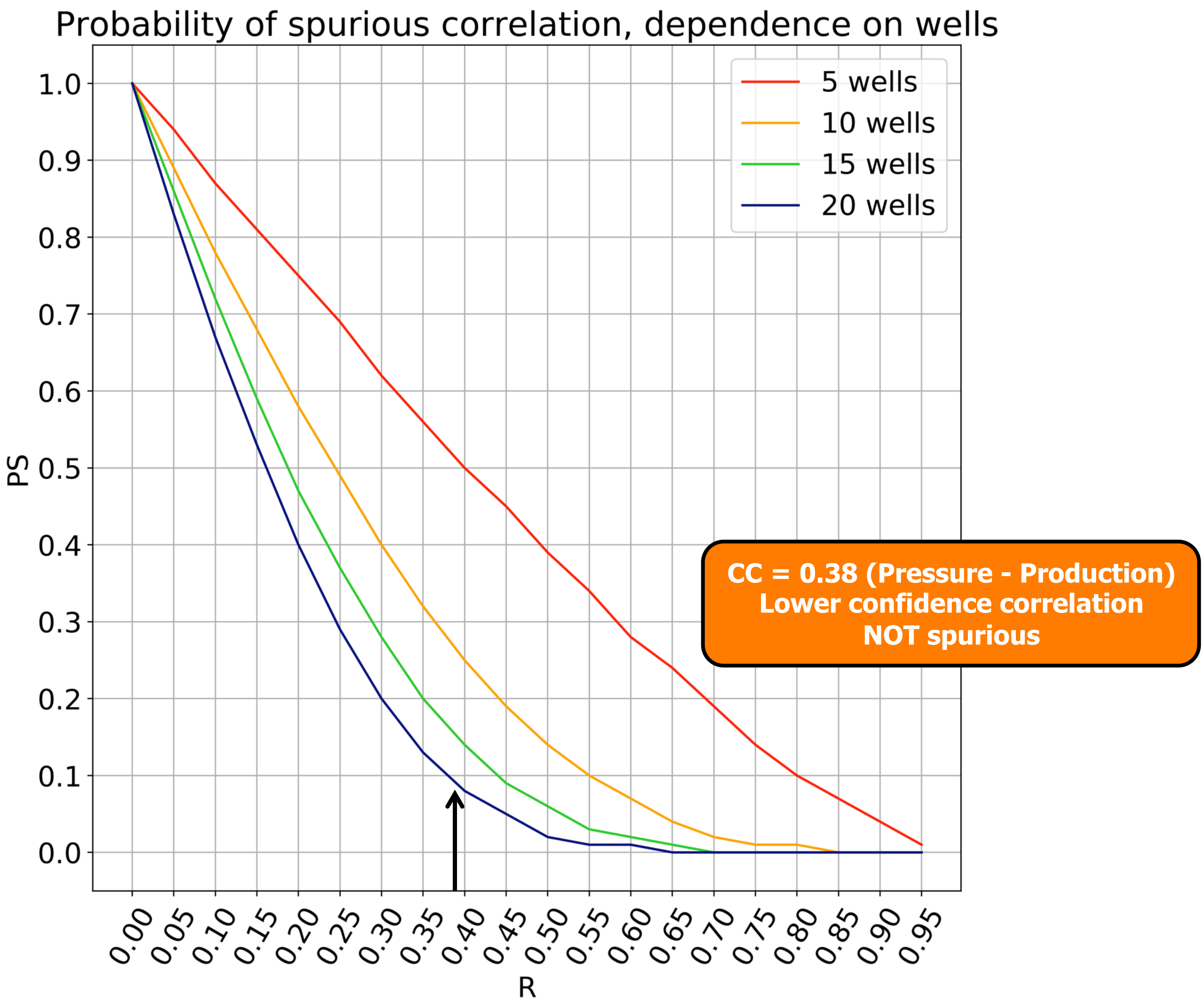
PS defined as the probability of observing the absolute value of the sample correlation, r, being greater than some constant, R, given the true (population) correlation ρ is zero

PS depends only on the number of wells n, the observed sample correlation (as compared to R), and the number of attributes

Probability of spurious correlation, 1 attribute

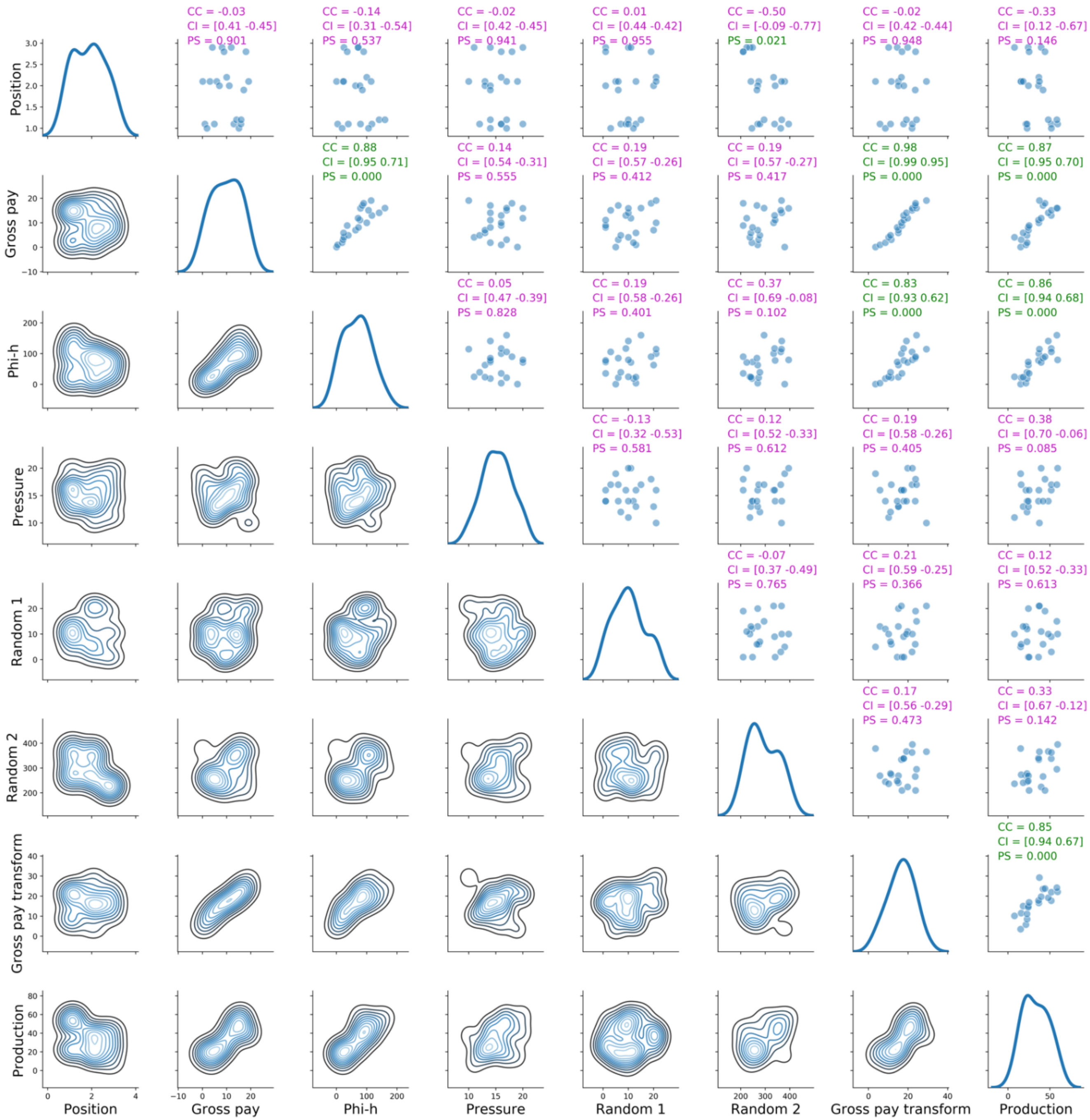
	n=5	n=10	n=15	n=20	n=25	n=35	n=50	n=75	n=100
R=0.1	0.87	0.78	0.72	0.67	0.63	0.57	0.49	0.39	0.32
R=0.2	0.75	0.58	0.47	0.4	0.34	0.25	0.16	0.09	0.05
R=0.3	0.62	0.4	0.28	0.2	0.15	0.08	0.03	0.01	0
R=0.4	0.5	0.25	0.14	0.08	0.05	0.02	0	0	0
R=0.5	0.39	0.14	0.06	0.02	0.01	0	0	0	0
R=0.6	0.28	0.07	0.02	0.01	0	0	0	0	0
R=0.7	0.19	0.02	0	0	0	0	0	0	0
R=0.8	0.1	0.01	0	0	0	0	0	0	0
R=0.9	0.04	0	0	0	0	0	0	0	0


Probability of spurious correlation, dependence on wells
5 wells
10 wells
15 wells
20 wells
PS
R
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
CC = 0.85 (Gross Pay - Production)
High confidence correlation
NOT spurious



Probability of spurious correlation, dependence on wells
5 wells
10 wells
15 wells
20 wells
PS
R
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
CC = 0.38 (Pressure - Production)
Lower confidence correlation
NOT spurious

When is a correlation statistically significant?



AGENDA

PART I (Python) – Statistically significant correlation

PART II (R)– Variable selection and multivariate analysis

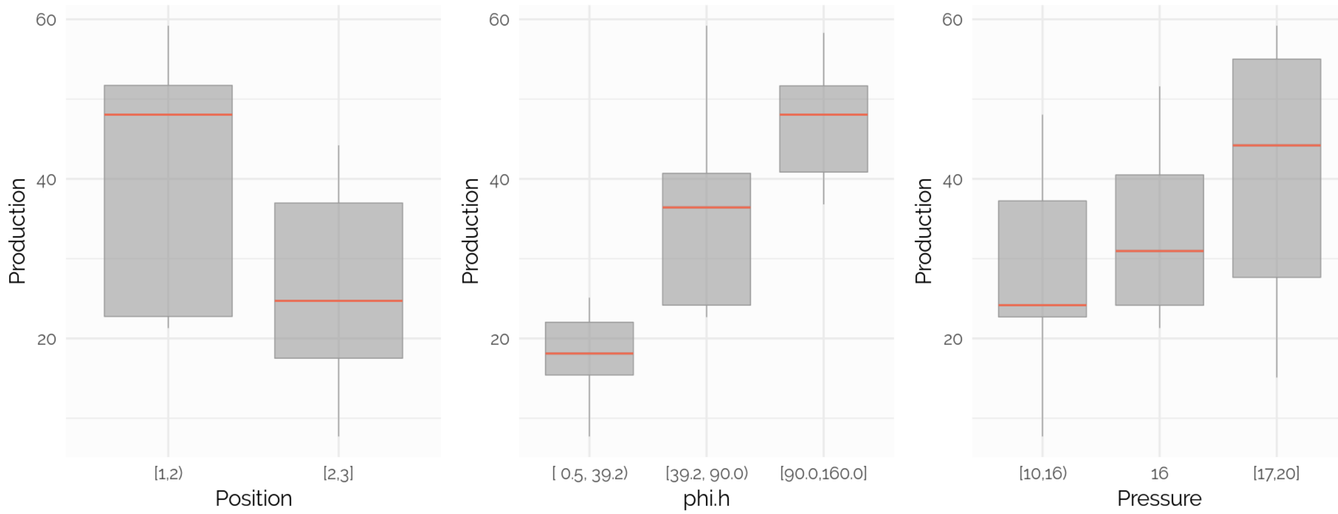
Correlation vs. Regression

- Correlation is the start of many analyses
- Has many similarities with regression
- Produces a score quantifying the strength of the **association** between pairs of variables

What if we want to understand these associations in more depth?

- On the **same unit** of analysis instead of a score
- Want to understand the **joint impact** of many variables on a response

Is there a difference in production due to …?



- ❑ Visual assessment via boxplot: makes us think about **uncertainty**!
- ❑ Can perform tests: is the difference in production statistically significant?
  E.g. test whether production **significantly different** according to position (Wilcoxon test)

estimate	statistic	p.value	conf.low	conf.high	method	alternative
14.98	83	0.041	4.64	27.3	Wilcoxon rank sum test	two.sided

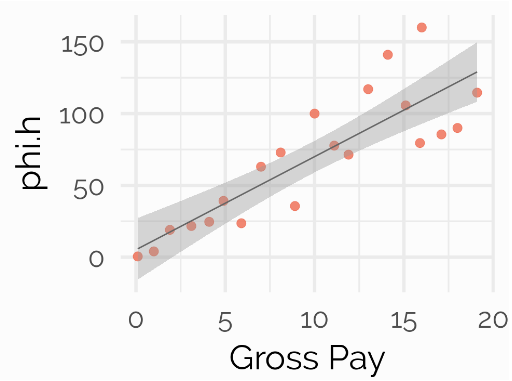
A P-value of **0.041**, tells us that the probability of observing a difference in distribution equal to or more extreme than the one observed is **4.1%.** In other words, it’s **somewhat unlikely** but not impossible that the difference in production is due to chance variation.

https://flowingdata.com/2008/02/15/how-to-read-and-use-a-box-and-whisker-plot/

Univariate Screening vs. Multivariable

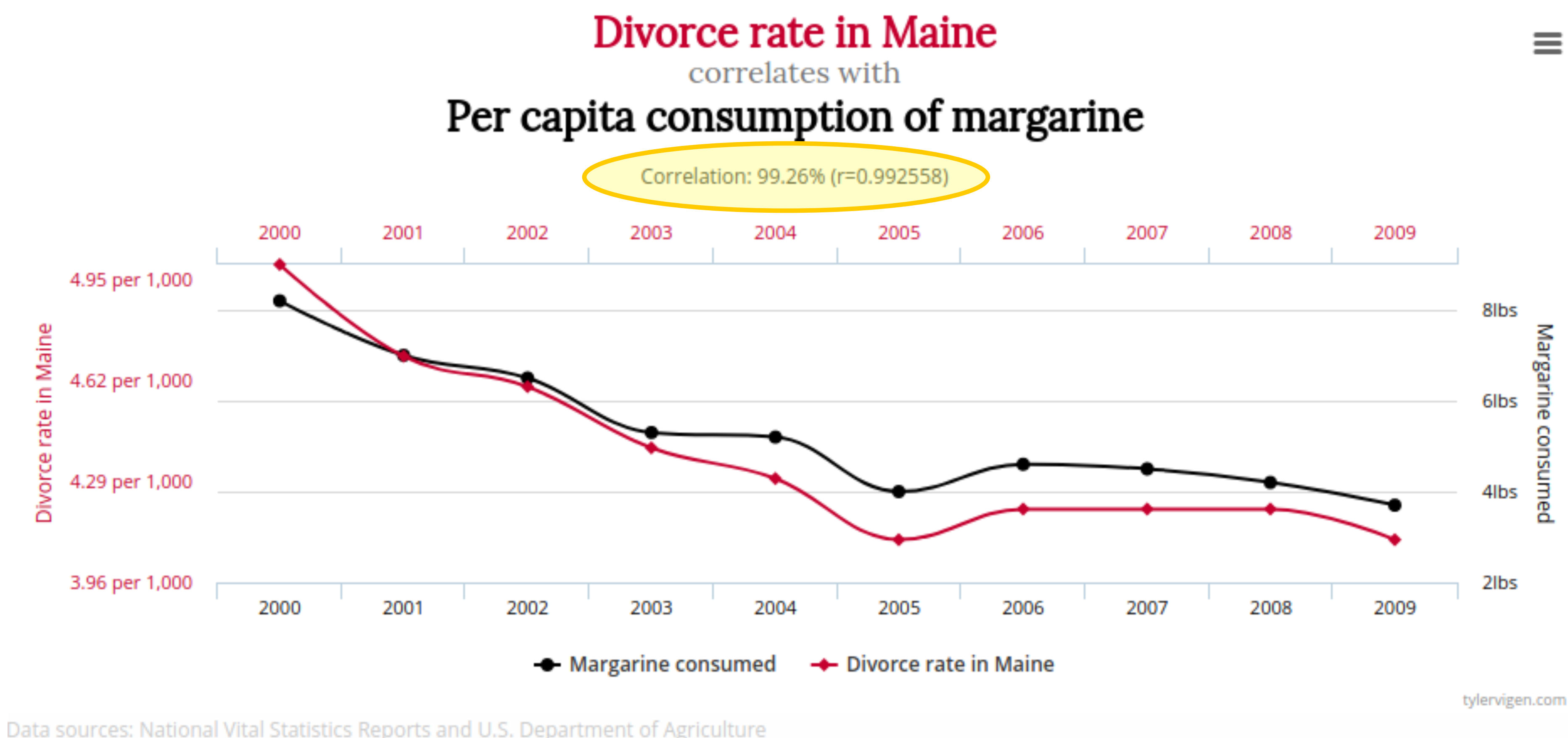
Method illustrated is called **univariate screening:** looks at each variable vs. response:

- ❑ Time consuming
- ❑ Cannot control for other variables. For instance: recall **gross pay** and **phi.h** are strongly related.
  - ❑ How do they **both** explain changes in production?
  - ❑ Can we get away with **just one** of them?
- ❑ Subject to **multiple comparison**: quite likely we will discover differences in production that are **not there**, simply due to random chance



Univariate Screening vs. Multivariable

What this talk is truly about… and what we want to avoid: **spuriousness**



Variable/Feature Selection

- Concerned with identifying a subset of "**important**" variables. Tied to the idea of parsimony and Occam's razor.

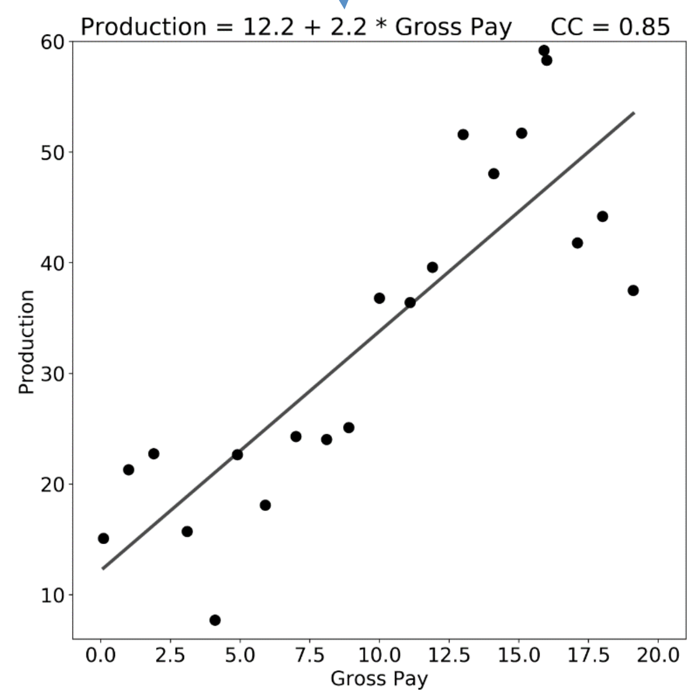
- Many methods, little consensus. Therefore, important to let **both** domain knowledge **and** statistical methods guide the process.

- We will illustrate variable selection via the **LASSO***.

** Many more methods are illustrated in the Github repo.*

Variable/Feature Selection via the LASSO

- **LASSO**: least absolute shrinkage and selection operator
- Recall from the previous slides a **regression model** for production
  - What if we were to **shrink** the coefficient for Gross Pay from 2.2 to 0?
  - Gross pay would no longer contribute to the model
  - We just achieved variable selection
  - LASSO shrinks coefficients via a regularization or shrinkage parameter λ, typically chosen via cross-validation
  - The cross-validated value minimizing the loss function is the proposed one



Variable/Feature Selection via the LASSO

Let's apply LASSO regression to our data. The response variable (Y) is production. The shrunk parameters are:

Variable	Shrinked Coefficient
(Intercept)	-4.278
gross.pay	1.032
phi.h	0.116
pressure	1.332
random.1	NA
random.2	0.010
gross.pay.transform	NA
position.cat	-6.544

Looking at the table of coefficients, notice how **random.1** and **gross.pay.transform** have no coefficient. That's because they have been shrunk to zero, thus achieving variable selection.

Our model suggests that these two variable are **not useful** in explaining changes in production.

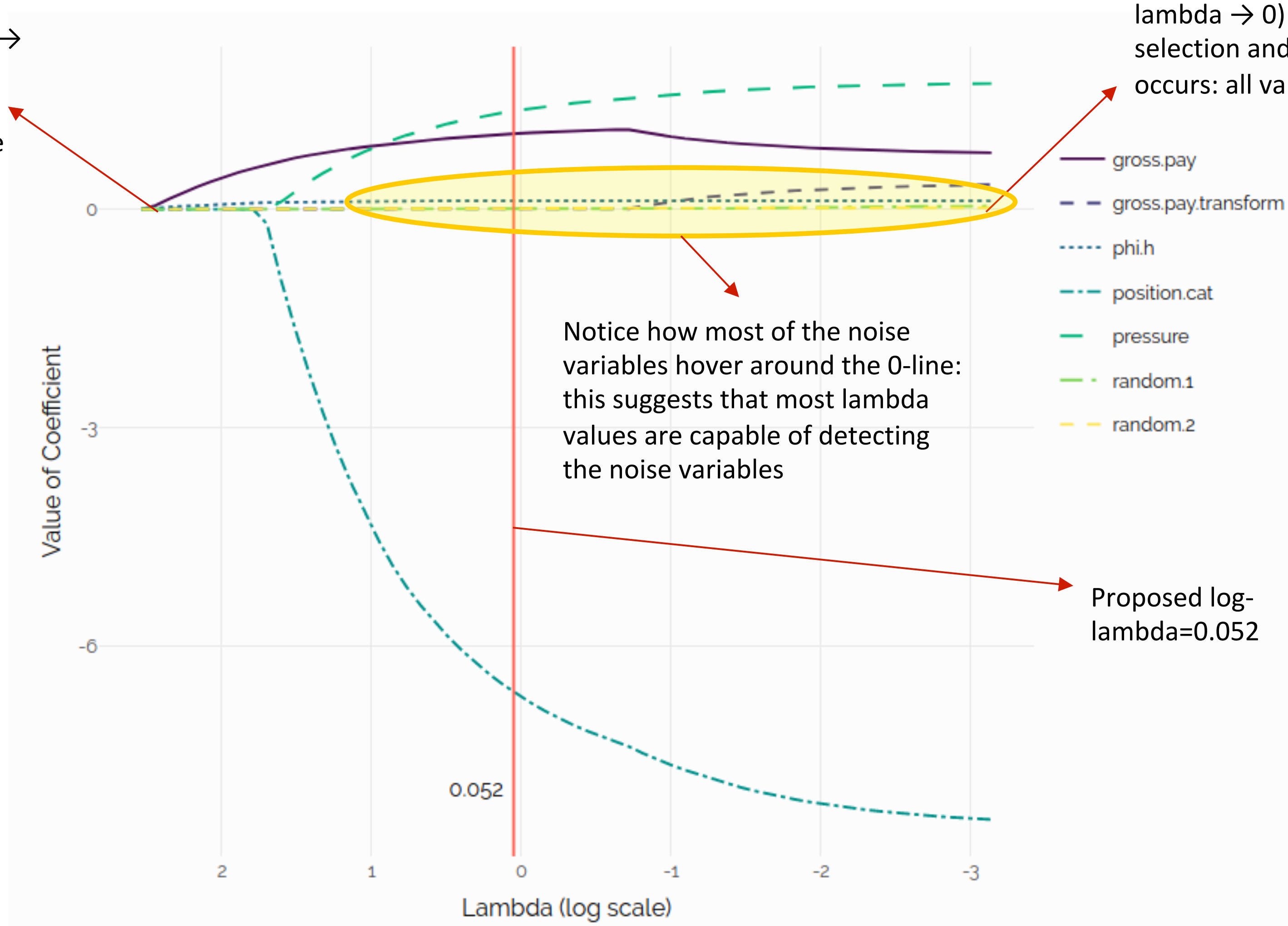
The coefficient for **random.2** is also nearly zero, so we could remove it as well.

Variable/Feature Selection via the LASSO

Let us see how "robust" the LASSO is in identifying these variables for different values of λ. The graph below shows how "quickly" variable are shrunk to zero (read from right to left)

As log-lambda → ∞, all variables are shrunk to 0 and no variable are selected

As log-lambda → - ∞ (i.e. lambda → 0) no variable selection and no shrinkage occurs: all var are retained



Using Regression to Understand Changes in Production

- Now that we have a better understanding of which variables are **"important"**, let's fit a regression model that **controls** for all known aspects of production
- We will use **phi.h**, **pressure**, and **position***

Modeling considerations

- With a sample size of **n=21**, **machine learning** approaches are "out for lunch"
- We will make **phi.h non-linear**
- Because our response, production, is a **rate**, we can probably use more appropriate methods than least squares
- Here we will use a type of semi-parametric model called **ordinal regression**
- Nice thing about ordinal regression is that **if** we had a reasonable sample size, we could **estimate any quantile of interest** (e.g. P25, P75)

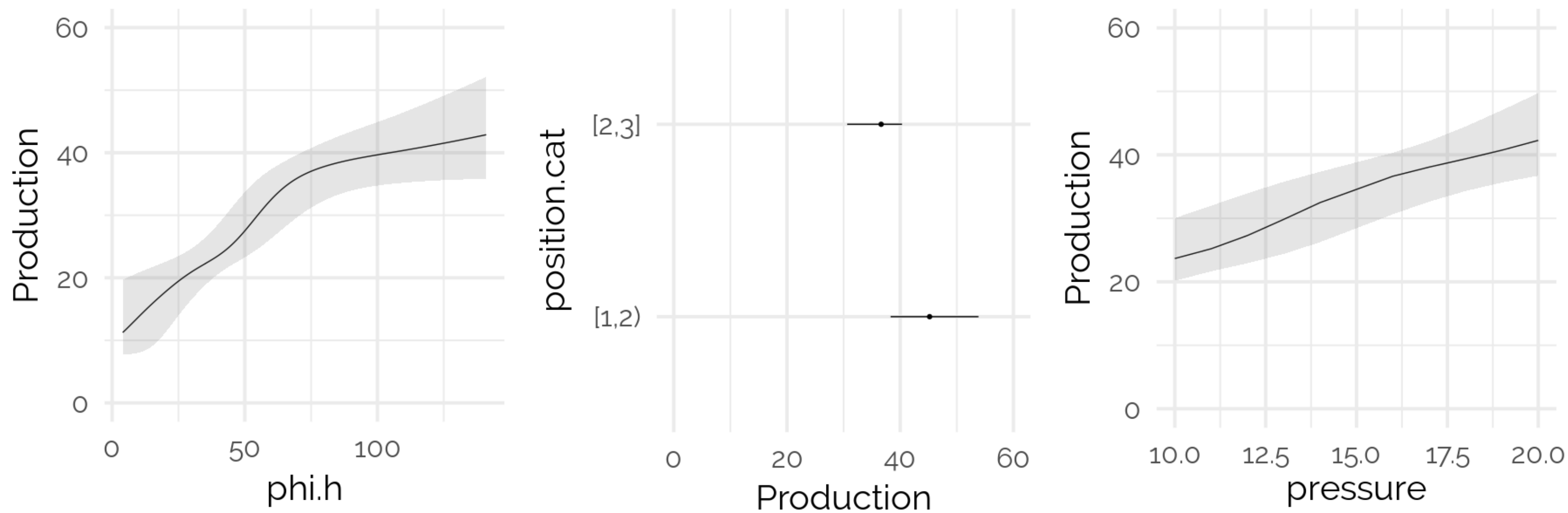
* See Github page for more information

Using Regression to Understand Changes in Production

	β	S.E.	Wald Z	Pr(>|Z|)
phi.h	0.1185	0.0253	4.68	<0.0001
phi.h'	-0.0515	0.0171	-3.02	0.0025
position.cat=[2,3]	-2.7903	0.7852	-3.55	0.0004
pressure	0.4958	0.1246	3.98	<0.0001

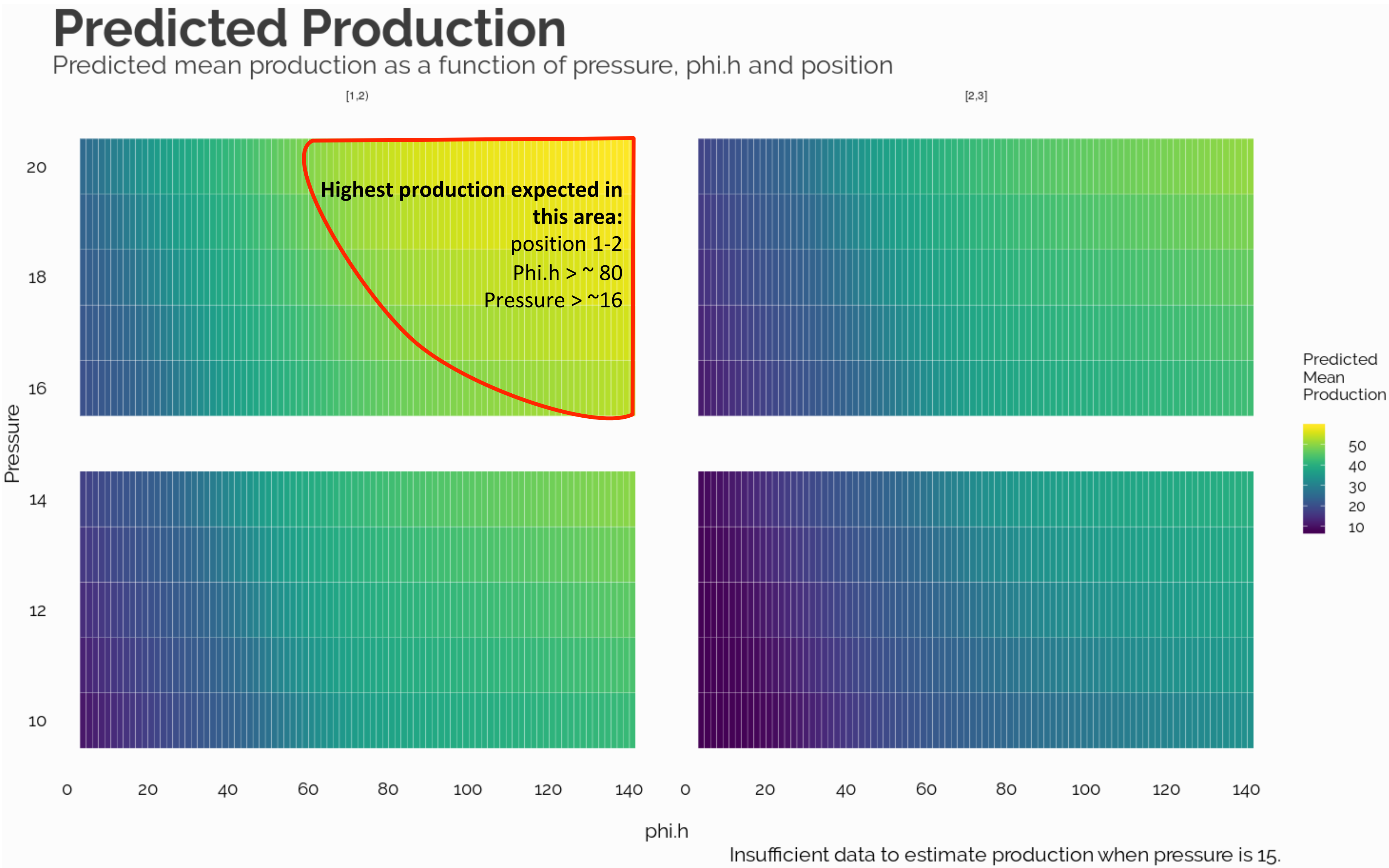
Table of coefficients

Easier to interpret graphically (below)



Predicted mean production vs. each variable while holding other variables fixed. In other words, effect of variable on production **over and above** any other variable

Using Regression to Understand Changes in Production



QUESTIONS ?

RESOURCES

Github repository:
https://github.com/mycarta/Niccoli_Speidel_2018_Geoconvention

REFERENCES

-Matteo Niccoli (Nov. 2016). Machine learning in geoscience with scikit-learn - notebook 2. GitHub notebook: github.com/mycarta/predict/blob/master/Geoscience_ML_notebook_2.ipynb

- Lee Hunt (Dec.2013). Many correlation coefficients, null hypoteses, and high value. Lee Hunt, CSEG Recorder.
- Cynthia Kalkomey (March 1997). Potential risks when using seismic attributes as predictors of reservoir properties., The Leading Edge.
- Richard Chambers and Jeffrey Yarus (June 2002). Quantitative use of seismic attributes for reservoir characterization., CSEG Recorder.
- Lee Hunt et al., (May 2014). Precise 3D seismic steering and production rates in the Wilrich tight gas sands of West Central Alberta. Lee Hunt et al., Interpretation.
- Stan Brown (updated 2018). Stats without tears. Free textbook at: brownmath.com/swt
- Frank E. Harrell, (2015) Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis. Second Edition. Springer-Verlag New York, Inc. New York, USA.