09 | Linear Regression, Part 2

Ivan Corneillet

Data Scientist

Learning Objectives

After this lesson, you should be able to:

- Define multiple linear regression
- Understand and identify multicollinearity in a multiple regression
- Evaluate a linear model fit's significance
- How to conduct linear regression modeling
- Explain the difference between causation and correlation

Multiple Linear Regression

Multiple Linear Regression

- Simple linear regression with one variable can explain some variance, but using multiple variables can be much more powerful
- We can extend this model to several input variables, giving us the multiple linear regression model

$$y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k + \varepsilon$$

- Given $x_j = \left(x_j^{(1)}, x_j^{(2)}, \dots, x_j^{(n)}\right)$, $y = \left(y^{(1)}, y^{(2)}, \dots, y^{(n)}\right)$, and $\varepsilon = \left(\varepsilon^{(1)}, \varepsilon^{(2)}, \dots, \varepsilon^{(n)}\right)$, we formulate the linear model as

$$y^{(i)} = \beta_0 + \beta_1 \cdot x_1^{(i)} + \cdots + \beta_k \cdot x_j^{(i)} + \varepsilon^{(i)}$$

- Given estimates for the model coefficients $\hat{\beta}_i$, we then predict y using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \cdots + \hat{\beta}_k \cdot x_k$$

Multiple Linear Regression (cont.)

- E.g. (SF housing dataset),

$$\widehat{SalePrice} = \hat{\beta}_0 + \hat{\beta}_1 \cdot Size + \hat{\beta}_2 \cdot LotSize$$

or

$$\widehat{SalePrice} = \hat{\beta}_0 + \hat{\beta}_1 \cdot Size + \hat{\beta}_2 \cdot Beds$$

Multiple Linear Regression

Common Regression Assumptions (cont.)

Common Regression Assumptions (Part 2)

- x_i are independent from each other (low multicollinearity)

- Multicollinearity (or collinearity) is a phenomenon in which two or more predictors in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy

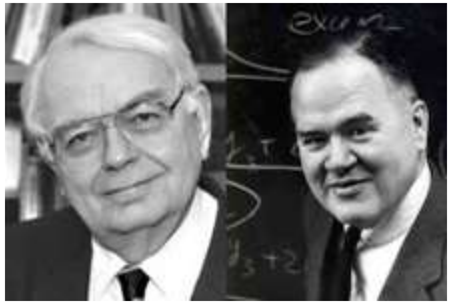
Ideal Scenario: When Predictors are Uncorrelated

- Each coefficient can be estimated and tested separately
- β_i estimates the expected change in y per unit change in x_i, all other predictors held fixed
- However predictors usually change together

- Correlations amongst predictors cause problems
  - The variance of all coefficients tends to increase, sometimes dramatically
  - Interpretations become hazardous – when x_i changes, everything else changes

The Woes of (Interpreting) Regression Coefficients

- “The only way to find out what will happen when a complex system is distributed is to disturb the system, not merely to observe it passively” – Fred Mosteller and John Tukey



- “Essentially, all models are wrong, but some are useful” – George Box

Common Regression Assumptions (Part 3)

- Linear regression also works best when
  - the data is normally distributed (it doesn't have to be)
    - (if data is not normally distributed, we could introduce *bias*)

`.plot_regress_exog()` (cont.)

- “Partial regression plot” (lower left)
- Partial regression for a single regressor
- The full model’s β_i is the fitted line’s slope
- The individual points can be used to assess the influence of points on the estimated coefficient
- `.plot_partregress()`

- “CCPR plot” (lower right)
  - Component and Component-Plus-Residual
- Refined partial residual plot
- Judge the effect of one regressor on the response variable by taking into account the effects of the other independent variables
- Scatterplot of the full model’s residuals ($\hat{\varepsilon}$) plus $\beta_i \cdot x_i$ against the regressor (x_i)
- `.plot_ccpr()`

Multiple Linear Regression

Assessing the Model's Fit with $\overline{\boldsymbol{R}}^{\mathbf{2}}$ *(adjusted* $\boldsymbol{R}^{\mathbf{2}}$*)*

Assessing the Model's Fit with $\bar{R}^2$ (adjusted R^2)

- R^2 increases as you add more variables in your model, even non-significant predictors; it's then tempting to add all the features from your dataset
- $\bar{R}^2$ attempts to adjust the explanatory power of regression models that contain different numbers of predictors so as to make comparisons possible

$$\bar{R}^2 = 1 - \left(1 - R^2\right) \cdot \frac{n-1}{n-k-1}$$

(n number of samples;
k number of parameters)

Linear Regression

Assessing the Model’s Fit Significance with the F-statistic

What β_i would make our multiple linear regression model useless?

- (the multiple linear regression model again)

$$y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k + \varepsilon$$

- Answer: If $\beta_0 = \beta_1 = \cdots = \beta_k = 0$, we don't have a linear model
  - ($y = 0$ isn't very exciting, is it?)

Model’s F-statistic Hypothesis Testing

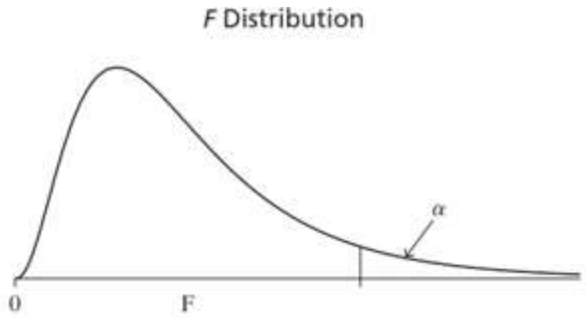
- The *null hypothesis* (H_0) represents the status quo; that the mean of all the regression’s coefficients β_i are equal to 0, i.e., that none of the β_i are significant:

$$H_0 \colon \forall i \colon \ \beta_i = 0$$

- The *alternate hypothesis* (H_a) represents the opposite of the null hypothesis (that at least one β_i is not zero) and holds true if H_0 is found to be false; that the mean of at least one the regression’s coefficientβ_i is not equal to 0, i.e. that at least one β_i is significant:

$$H_a \colon \exists i \colon \ \beta_i \neq 0$$

The F-distribution table ($\alpha = .05$) ($df_1 = k, df_2 = n$)



$\alpha = .05$

df_2 \ df_1	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	[illegible]	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00

Linear Regression

How to conduct Linear Regression Modeling

and Stepwise Model Selection Procedures

How to conduct Linear Regression Modeling

❶ Model's Significance

- Always start with the F-statistics for the whole model; only then check individual features

❷ Regressors' Significance

- Prefer to work solely with significant features: if you observe insignificant features you *usually* need to get rid of them and rerun your regression modeling without those

Stepwise model selection procedures

Forward Selection

- Start with no feature in the model and add features to the model one at a time. At each step, each feature not already in the model is tested for inclusion. The most significant of these features (if any). Repeat this process until no additional feature improves the model to a statistically significant extent
- However, the addition of a new feature may render one or more of the already included variables non-significant; backward selection avoids this drawback

Backward Selection

- Start from the “other end” by fitting a model with all the features of interest. If you have insignificant features, start dropping the most insignificant feature. If after removing that feature you still have insignificant features, drop them one by one, until you are left with no insignificant feature
- Sometimes dropped features would become significant if added to the final reduced models. Compromise between forward and backward selection methods should be considered, e.g., bidirectional elimination, testing at each step for features to be included or excluded

Problems with stepwise model selection procedures

- “... perhaps the most serious source of error lies in letting statistical procedures make decisions for you”
- “Don't be too quick to turn on the computer. Bypassing the brain to compute by reflex is a sure recipe for disaster”
  - – Phillip Good and James Hardin, Common Errors in Statistics (and How to Avoid Them)

Data Mining

Why is this?

- Sensational headlines
- No robust data analysis
- Lack of understanding of the difference between *causation* and *correlation*
  - **“caused” ≠ “measured” or “associated”**
  - ***Correlation does not imply causation***

- Understanding this difference is critical in the data science workflow, especially when **Identifying** the problem and **Acquiring** the data
  - We need to fully articulate our question and use the right data to answer it, including any *confounders*
- Additionally, this comes up when **Presenting** our results to stakeholders

Do you really need causality or is correlation enough?

Collaborative recommendations using item-to-item similarity mappings

US 6266649 B1

Publication number	US6266649 B1
Publication type	Grant
Application number	US 09/157,198
Publication date	Jul 24, 2001
Filing date	Sep 18, 1998
Priority date	Sep 18, 1998
Fee status	Paid
Also published as	EP1121658A1, EP1121658A4, WO2000017792A1
Inventors	Gregory D. Linden, Jennifer A. Jacobi, Eric A. Benson
Original Assignee	Amazon.Com, Inc.
Export Citation	BiBTeX, EndNote, RefMan

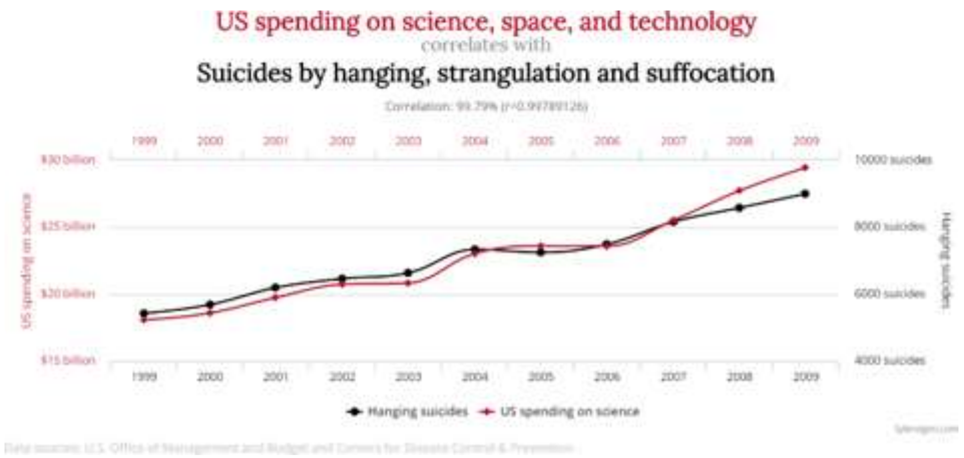
Patent Citations (22), Non-Patent Citations (39), Referenced by (1104), Classifications (23), Legal Events (9)

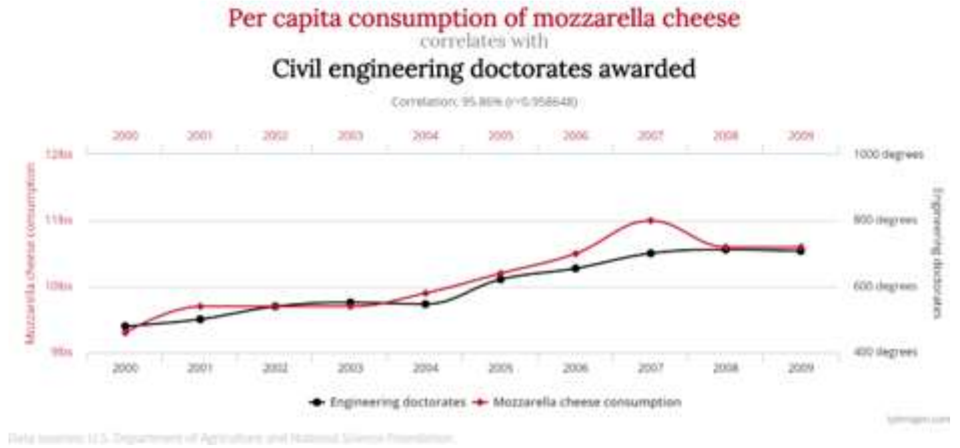
External Links: USPTO, USPTO Assignment, Espacenet

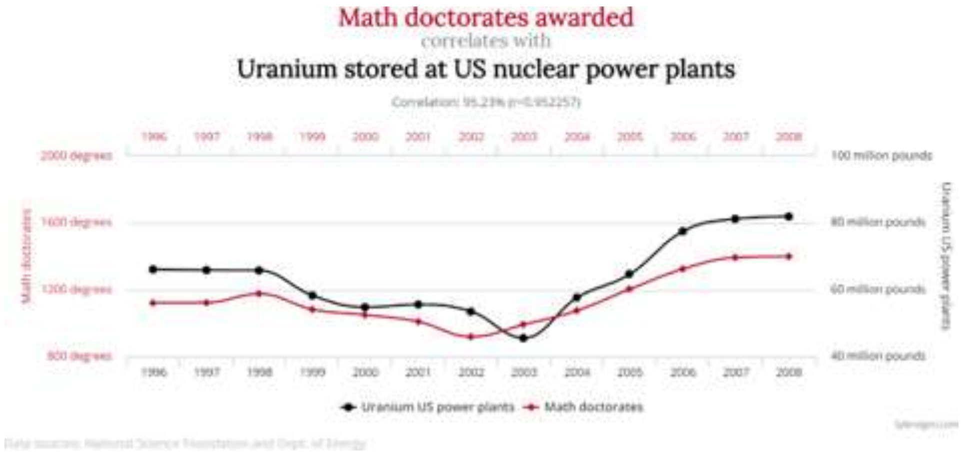
ABSTRACT

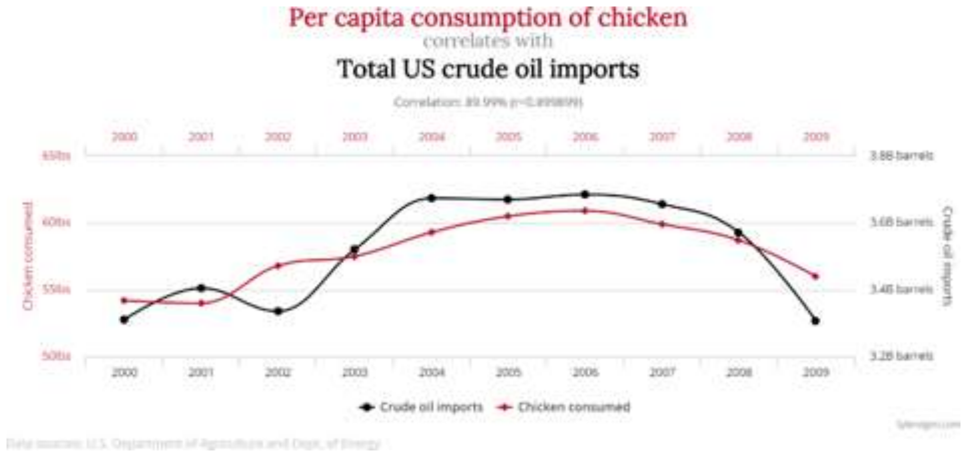
A recommendations service recommends items to individual users based on a set of items that are known to be of interest to the user, such as a set of items previously purchased by the user. In the disclosed embodiments, the service is used to recommend products to users of a merchant's Web site. The service generates the recommendations using a previously-generated table which maps items to lists of "similar" items. The similarities reflected by the table are based on the collective interests of the community of users. For example, in one embodiment, the similarities are based on correlations between the purchases of items by users (e.g., items A and B are similar because a relatively large portion of the users that purchased item A also bought item B). The table also includes scores which indicate degrees of similarity between individual items. To generate personal recommendations, the service retrieves from the table the similar items lists corresponding to the items known to be of interest to the user. These similar items lists are appropriately combined into a single list, which is then sorted (based on combined similarity scores) and filtered to generate a list of recommended items. Also disclosed are various methods for using the current and/or past contents of a user's electronic shopping cart to generate recommendations. In one embodiment, the user can create multiple shopping carts, and can use the recommendation service to obtain recommendations that are specific to a designated shopping cart. In another embodiment, the recommendations are generated based on the current contents of a user's shopping cart, so that the recommendations tend to correspond to the current shopping task being performed by the user.

Spurious Correlations









Source: tylervigen.com

Linear Regression

statsmodels vs. sklearn

statsmodels vs. *sklearn*

	Pros	Cons
statsmodels *(Takeaway: Use statsmodel for your modelling's inner-loop)*	❑ Does linear regression modelling very well ❑ Very convenient summary report about your model's fit: mode's F-value/p-value; model's coefficients t-values, p-values, and confidence intervals ❑ Enable for quick iterations during exploratory data analysis and modeling phases	❑ Limited to a few types of models
sklearn *(Takeaway: Use sklearn to validate your model and then afterwards for production/prediction purpose)*	❑ Consistent programming interface to build many different types of machine learning models ❑ Facilities to validate models' fit (i.e., validation, cross-validation, …)	❑ Doesn't provide an easy-to-read summary report for your linear regression model. E.g., no F-value for the entire model is reported

Linear Regression

Pros and Cons

Linear Regression | Pros and cons

- Pros
  - Intuitive, well-understood, highly interpretable, and simple to explain
  - Can perform well with a small number of samples
  - Model training and prediction are fast
  - No need to standardize your data (i.e., features don't need scaling)
  - No tuning is required (excluding regularization)

- Cons
  - Assumes linear association among variables
  - Assumes normally distributed residuals
  - Outliers can easily affect coefficients