Linear Regression - Predictive and Inferential

Objectives

After this lecture you should be able to:

- Distinguish between predictive and inferential linear regression
- Be able to define and detect collinearity between features
- Be able to encode binary and categorical features

Predictive Linear Regression

Goal:
Accurately predict a target

We picked a model based on trying different features, feature engineering, evaluation using cross validation.

We care that it predicts well on unseen data.

We don’t really care that:

- Some of the features may be partially collinear (more later)
- Because of that, we can’t rely on our parameter estimates to tell us something about their effect on the signal
- We may be violating some fundamental assumptions of inferential linear regression (more later)

Inferential Linear Regression

Goal:
Learn something accurate about the process that made the data. Infer (estimate) coefficients.

We picked a model based on trying different features, feature engineering, and checking residuals to see if we are violating some of the assumptions of linear regression.

We care that the parameter estimates are accurate and valid.

We don't really care that:

- It predicts well (but, it should!)

So what should you do?

It depends. Most often, we're asked for predictive models.

BUT:
One benefit of linear regression is interpretability of the coefficients. If assumptions of (inferential) linear regression are being violated, then you can't rely on your parameter estimates to provide that interpretability.

So often a hybrid approach is taken: predictive (best model through cross-validation) but attempt to detect and remove collinearity, verify normal distribution of errors, ensure data points are independent (more later) so that you have confidence in your parameter estimates.

Have two notebooks we'll go through at the end:
`predictive-linear-regression.ipynb`
`inferential-linear-regression.ipynb`

Collinearity

Multicollinearity (also collinearity) is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy.

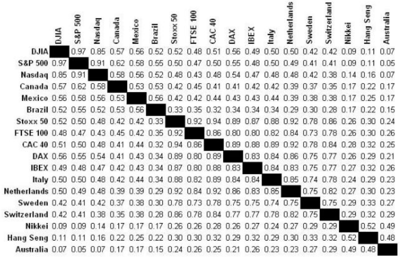
In this situation the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data.

Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set; it only affects calculations regarding individual predictors.

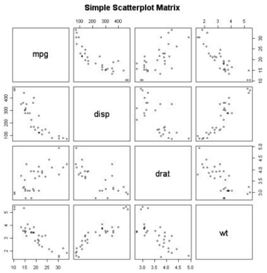
source: Wikipedia

Detecting collinearity

- Correlation Matrix / Scatterplot Matrix



	DJIA	S&P 500	Nasdaq	Canada	Mexico	Brazil	Stoxx 50	FTSE 100	CAC 40	DAX	IBEX	Italy	Netherlands	Sweden	Switzerland	Nikkei	Hang Seng	Australia
DJIA		0.97	0.85	0.57	0.56	0.52	0.52	0.48	0.51	0.56	0.49	0.50	0.50	0.42	0.42	0.09	0.11	0.07
S&P 500	0.97		0.91	0.62	0.58	0.55	0.50	0.47	0.50	0.55	0.48	0.50	0.49	0.41	0.41	0.09	0.11	0.05
Nasdaq	0.85	0.91		0.58	0.56	0.52	0.48	0.43	0.48	0.54	0.47	0.48	0.48	0.42	0.38	0.14	0.16	0.07
Canada	0.57	0.62	0.58		0.53	0.53	0.42	0.45	0.41	0.41	0.42	0.42	0.39	0.37	0.35	0.17	0.22	0.17
Mexico	0.56	0.58	0.56	0.53		0.56	0.42	0.42	0.44	0.43	0.43	0.44	0.39	0.38	0.38	0.17	0.25	0.17
Brazil	0.52	0.55	0.52	0.53	0.56		0.33	0.35	0.32	0.34	0.34	0.34	0.29	0.30	0.28	0.17	0.22	0.15
Stoxx 50	0.52	0.50	0.48	0.42	0.42	0.33		0.92	0.94	0.89	0.87	0.88	0.92	0.78	0.86	0.26	0.30	0.24
FTSE 100	0.48	0.47	0.43	0.45	0.42	0.35	0.92		0.86	0.80	0.80	0.82	0.84	0.73	0.78	0.26	0.30	0.26
CAC 40	0.51	0.50	0.48	0.41	0.44	0.32	0.94	0.86		0.89	0.88	0.89	0.92	0.78	0.84	0.28	0.32	0.25
DAX	0.56	0.55	0.54	0.41	0.43	0.34	0.89	0.80	0.89		0.83	0.84	0.86	0.75	0.77	0.26	0.29	0.21
IBEX	0.49	0.48	0.47	0.42	0.43	0.34	0.87	0.80	0.88	0.83		0.84	0.83	0.75	0.77	0.27	0.32	0.26
Italy	0.50	0.50	0.48	0.42	0.44	0.34	0.88	0.82	0.89	0.84	0.84		0.85	0.74	0.78	0.24	0.29	0.23
Netherlands	0.50	0.49	0.48	0.39	0.39	0.29	0.92	0.84	0.92	0.86	0.83	0.85		0.75	0.82	0.27	0.30	0.23
Sweden	0.42	0.41	0.42	0.37	0.38	0.30	0.78	0.73	0.78	0.75	0.75	0.74	0.75		0.75	0.29	0.33	0.27
Switzerland	0.42	0.41	0.38	0.35	0.38	0.28	0.86	0.78	0.84	0.77	0.77	0.78	0.82	0.75		0.29	0.32	0.29
Nikkei	0.09	0.09	0.14	0.17	0.17	0.17	0.26	0.26	0.28	0.26	0.27	0.24	0.27	0.29	0.29		0.52	0.49
Hang Seng	0.11	0.11	0.16	0.22	0.25	0.22	0.30	0.30	0.32	0.29	0.32	0.29	0.30	0.33	0.32	0.52		0.48
Australia	0.07	0.05	0.07	0.17	0.17	0.15	0.24	0.26	0.25	0.21	0.26	0.23	0.23	0.27	0.29	0.49	0.48	



Downside is can only pick up pairwise effects ☹

If there there is feature that's collinear with other features, remove it!

- Variance Inflation Factors (VIF)
  - Run ordinary least squares for each predictor as function of all the other predictors. k times for k predictors

$$X_1 = \alpha_2 X_2 + \alpha_3 X_3 + \cdots + \alpha_k X_k + c_0 + e$$

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

Looks at all predictors together! ☺

Rule of Thumb, > 10 is problematic

[Stats Models Variance Inflation Factor](#)

Handling categorical features

Binary encoding (yes/no, true/false, exists or not).
Example: Sex (Male/Female)

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

$$y_i = \beta_0 + \beta_1 \underline{x_i} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

B_1 quantifies how much being female changes the response relative to male.

Handling categorical features

Multi-category encoding
Example: Ethnicity (Asian/Caucasian/African American)

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}$$

$$y_i = \beta_0 + \beta_1 \underline{x_{i1}} + \beta_2 \underline{x_{i2}} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

B_1 quantifies how much being Caucasian changes the response relative to African American.

Data

Ones	Ethnicity
1	AA
1	Asian

Recode Design Matrix

Ones	Asian	Caucasian
1	0	0
1	1	0

Objectives

After this lecture you should be able to:

- Distinguish between predictive and inferential linear regression
- Be able to define and detect collinearity between features
- Be able to encode binary and categorical features

`predictive-linear-regression.ipynb`

`inferential-linear-regression.ipynb`