Business Analytics

Session 2a. More on Linear Regression

Renyu (Philip) Zhang

New York University Shanghai

Spring 2019

Warm-up Exercises

- What is the purpose of doing supervised learning?
  - Making accurate predictions about the outcome variable(s).

- What is the difference between supervised and unsupervised learning?
  - Supervised learning seeks to predict the outcome variable whereas unsupervised learning aims to find the structures in the data.

- How do we interpret R^2 in linear regression?
  - The ability of fitted/predicted outcome values to "explain" the sample variability of the true outcome values.

Wine Analytics

R^2 for Models with More Covariates

Covariates	R^2
AGST	0.44
AGST+Harvest Rain	0.71
AGST+Harvest Rain+Age	0.79
AGST+Harvest Rain+Age+Winter Rain	0.83
AGST+Harvest Rain+Age+Winter Rain+Population	0.83

- Adding more variables improves R^2 with diminishing returns.

R^2 for Models with More Covariates

Covariates	R^2
AGST	0.44
AGST+Harvest Rain	0.71
AGST+Harvest Rain+Age	0.79
AGST+Harvest Rain+Age+Winter Rain	0.83
AGST+Harvest Rain+Age+Winter Rain+Population	0.83

- Adding more variables improves R^2 with diminishing returns.
- Not all variables should be used.
  - More variables require more data.
  - Over-fitting (bad performance on unseen data).
  - How to choose variables to remove? Significance and correlations.

Understanding the Model

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.504e-01  1.019e+01  -0.044 0.965202
AGST         6.012e-01  1.030e-01   5.836 1.27e-05 ***
HarvestRain -3.958e-03  8.751e-04  -4.523 0.000233 ***
WinterRain   1.043e-03  5.310e-04   1.963 0.064416 .
Age          5.847e-04  7.900e-02   0.007 0.994172
FrancePop   -4.953e-05  1.667e-04  -0.297 0.769578
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Significance ($p-$value): Which covariates have (statistically) significant association with the outcome?
  - Consider removing insignificant covariates from the model (Age and France Population).
- Sign of coefficients: Positive or negative association with the outcome?
- Estimated values of coefficients: How strong are the associations?

Correlations between Covariates

- Correlation: Meassure of dependence/association between two covariates X and Z.

$$\rho_{XZ} = \frac{\text{Cov}(X, Z)}{\sigma_X \sigma_Z} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}))(Z_i - \bar{Z})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2]}\sqrt{\sum_{i=1}^{n}(Z_i - \bar{Z})^2}} \in [-1, 1]$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } \bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

  - $\rho_{XZ} > 0$: Positive relationship.
  - $\rho_{XZ} = 1$: Perfect positive relationship.
  - $\rho_{XZ} = 0$: "No" relationship.
  - $\rho_{XZ} < 0$: Negative relationship.
  - $\rho_{XZ} = -1$: Perfect negative relationship.
- Consider removing one of the two covariates who have strong positive or negative correlation.
  - The correlation between Age and France Population is -0.994.
  - Remove France Population from the covariates.

Predictive Power

- Out linear regression model has $R^2 = 0.83$.
  - In-sample fit between model and training data.
- How does the model perform on a new testing data set?
  - Out-of-sample R^2: $1 - SSE/SST$ on a new data set.

Covariates	R^2	Out-of-Sample R^2
AGST	0.44	0.79
AGST+Harvest Rain	0.71	-0.08
AGST+Harvest Rain+Age	0.79	0.53
AGST+Harvest Rain+Age+Winter Rain	0.83	0.79
AGST+Harvest Rain+Age+Winter Rain+Population	0.83	0.76

- Better in-sample R^2 does not necessarily mean better testing R^2.
- Out-of-sample R^2 can be negative.
- Any issue with our testing?

Model Interpretations and Recommendations

- lm(Price∼AGST+Harvest Rain+Age+Winter Rain)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.4299802  1.7658975  -1.942 0.066311 .
AGST         0.6072093  0.0987022   6.152  5.2e-06 ***
HarvestRain -0.0039715  0.0008538  -4.652 0.000154 ***
WinterRain   0.0010755  0.0005073   2.120 0.046694 *
Age          0.0239308  0.0080969   2.956 0.007819 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- How do we interpret the results?
  - Correlation/association?
  - Prediction?
  - Causation?
- What recommendations can we make?
  - Store the wine produced with high AGST, low harvest rain, and high winter rain.

Analytics vs. Expert

- Expert
  - 1986 is "very good to sometimes exceptional".
- Analytics
  - 1986 is medicore.
  - 1989 will be "the wine of the century" and 1990 will be even better.
- In wine auctions (the real market)
  - 1989 sold for more than twice the price of 1986.
  - 1990 sold for even higher prices.

More on Linear Regression

Interpreting Regression Coefficients

$$Y_i \approx \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$$

- How to interpret the coefficients?
  - $\hat{\beta}_0$ is the fitted outcome value when the covariates $X_i = (0, 0, \cdots, 0)$.
  - $\hat{\beta}_j$ is the change in the fitted outcome value for a unit change in the j'th covariate, *with all other covariates constant*.
- When we can/cannot say the following:
  (a) A unit change in X_{ij} is associated/correlated with a $\hat{\beta}_j$ unit change in Y_i.
  (b) Given $(X_{i1}, X_{i2}, \cdots, X_{ip})$, we predict Y_i will be $\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$.
  (c) A unit change in X_{ij} leads to a $\hat{\beta}_j$ unit change in Y_i.
- If we blindly run a linear regression, (a) is valid, but not (b) and (c).
  - We will address (b) and (c) later in this course.

Collinearity and Identifiability

- If $\hat{\beta}$ is not unique under OLS, the model is called nonidentifiable.
  - This occurs when one covariate can be expressed as the linear combination of other covariates.
  - This problem is referred to as collinearity and the resulting model is nonidentifiable.

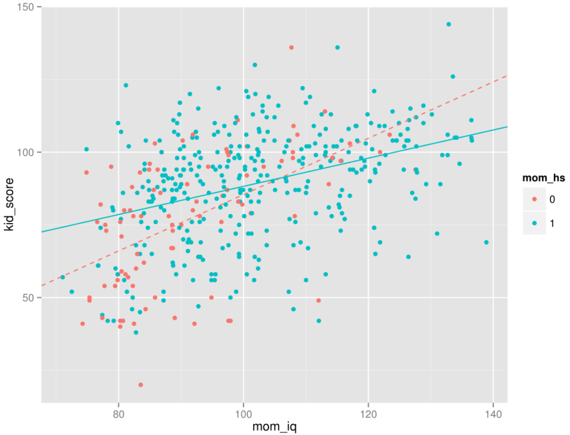
- Example: We add a new coefficient to the wine analytics model, *Year*, which refers to the year the wine was produced.

- If $p \approx n$ and n is large, this is called the high-dimensional regime.
  - If $p + 1 \geq n$, a linear regression model can perfectly fit the data. Is this a good model?
  - If $p \geq n$, the model is nonidentifiable.

Beyond Linearity

- What if the relationship between Y and X are inherently non-linear?
  - Include higher order terms, $(X_i)^2$, or cross terms $X_j X_k$.
  - The context and domain knowledge are important.
- If the value of one covariate affects the slope of another, then we need an interaction term.
  - Demo: Child IQ.
- Rules-of-thumb
  - Include interactions with a covariate with large coefficient in a standard linear regression.
  - Include interactions with a covariate describing groups of data (e.g., Child IQ example).

Visualization of Interactions



kid_score $\approx$
$-11.48 + 0.97$*mom_iq* if
mom_hs $= 0$

kid_score $\approx$
$39.79 + 0.48$*mom_iq* if
mom_hs $= 1$

Data Transformation

- In a lot of cases, working with transformed versions of the data makes the regression more meaningful.

- Two commonly used transformations:
  - Logarithms of positive variables
  - Centering and standardizing

Logarithmic Transformation

- In some contexts, the outcome Y is positive (e.g., height, counts, prices, revenues, etc.).
- The linear regression model may be problematic, because Y_i may be negative for some covariate X_i.
- One approach is to take logarithmic transformation of the data before applying linear regression (e.g., the log of wine price).

$$\log(Y_i) \approx \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{ij}$$

- One unit change in X_{ij} is suggesting a proportional change of fitted outcome by $\exp(\hat{\beta}_j) \approx 1 + \hat{\beta}_j$.
  - If both the outcome and the covariates are logged, the coefficients give proportional changes in the fitted outcome associated with the proportional changes in covariates.

Centering

- Sometimes to make coefficients interpretable, it is useful to center covariates by removing the mean $\tilde{X}_{ij} = X_{ij} - \bar{X}_j$, where $\bar{X}_i = \frac{1}{n}\sum_i X_{ij}$.
- $Y_i \approx \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j \tilde{X}_{ij}$.
- Example: Wine Analytics. Any observations? How to interpret $\hat{\beta}_0$?
  - The intercept can be directly interpretable as the average value of the outcome when the covariates are around the average.
- Sometimes it is useful to standardize $\tilde{X}_{ij} = \frac{X_{ij} - \bar{X}_j}{\hat{\sigma}_j}$
  - How to interpret $\hat{\beta}_j$ in this case?