Parametric Statistics

Week 10 - Multiple Linear Regression

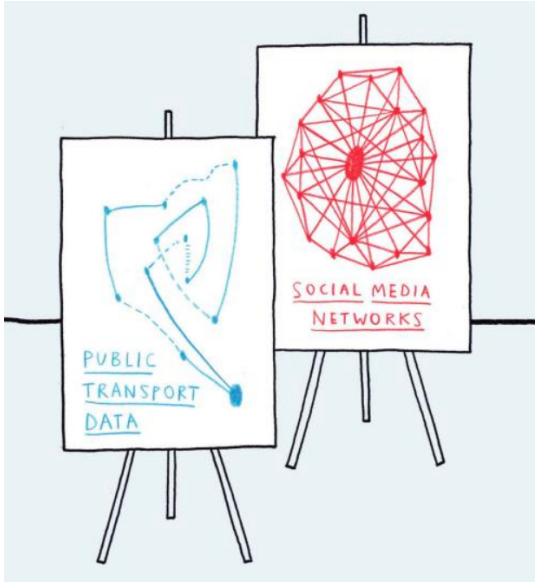
Juan Carlos Medina Serrano

Technische Universität München
Hochschule für Politik
Political Data Science

Munich, 15. January 2020

political data science

https://politicaldatascience.blogspot.de



Multiple Linear Regression

Now we will have more than one predictor. Each predictor will have its own regression coefficient:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + error$$

We predict Y:

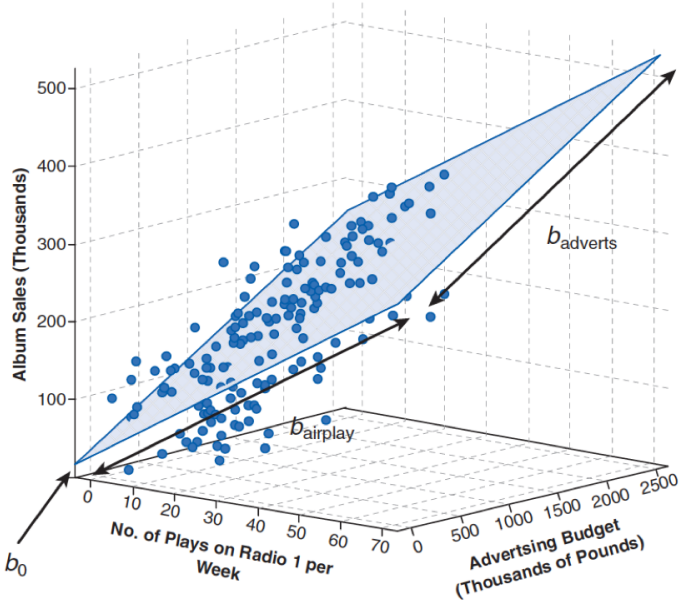
$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

Note The order of the βs and Xs is not important, only β_0 is without predictor. Think better that every predictor has a regression coefficient.

Multiple Linear Regression

Example Predict Album sales from the advertising budget and the number of times the songs were played on the radio

$$Album\ Sales = \beta_0 + \beta_1 * Advertising + \beta_2 * RadioPlays$$



In the figure, the $b_0, b_{airplay}, b_{adverts}$ correspond to $\beta_0, \beta_1, \beta_2$

Multiple Linear Regression

How to find the best linear model? = How to find the best βs?

With the coefficients that minimize the errors. (squared errors). We use again the *residual sum of squares* as:

$$RSS = error_1^2 + error_2^2 + ... + error_N^2$$

or equivalent:

$$RSS = (y_1 - \beta_0 - \beta_1 x1_1 - \beta_2 x2_1 - ...)^2 + (y_2 - \beta_0 - \beta_1 x1_2 - \beta_2 x2_2 - ...)^2 + ... + (y_N - \beta_0 - \beta_1 x1_N - \beta_2 x2_N - ...)^2$$

By minimizing the RSS, we obtain the coefficients. The math requires knowledge of linear algebra. We will skip it for this course.

How good is the model?

In the same way as with linear regression, we can calculate the $SS_{total}, SS_{model}, SS_{residual}$.

We also use the R^2 measure of fit. However, it is a multiple R^2, which is the square of the correlation between the real Y and the predicted Y^2:

$$R^2 = corr(Y, \hat{Y})^2$$

It represents **the amount of variation of Y that can be explained by the model.**

How good is the model?

How significant is the model?

Remember: For linear regression, we used the ANOVA F-test of the complete model and the t-test for individual coefficients.

For multiple linear regression, we **first** use the F-test to find if the complete model is significant:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

$$H_1 : At\ least\ one\ \beta\ is\ not\ zero$$

F-statistic is the same as for linear regression:

$$F = \frac{variance\ explained}{variance\ not\ explained} = \frac{MS_{model}}{MS_{residuals}}$$

The higher the F value, the more statistical significance.

How good is the model?

Only after we know that the F-test is significant, we can continue to test which predictors are significant.

We achieve this with t-tests for each coefficient.

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

The t-test statistic:

$$t = \frac{\beta_j - 0}{SE(\beta_j)}$$

The standard errors can help us to build also confidence intervals for each coefficient.

Model Output

Example We want to predict the number of votes for local politicians with the help of advertising in TV, radio and newspapers:

$$Votes = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper$$

Results:

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	-0.001	0.0059	-0.18	0.8599

Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistic	570

Interpretation: The model is significant because of the large F value. Moreover it explains 89.7% of the variance in Votes. TV and radio are significant, but Newspaper is not. Radio has a larger coefficient than TV, so it means that radio advertising helped more than TV. What is the residual standard error?

Model Output

Example We want to predict the number of votes for local politicians with the help of advertising in TV, radio and newspapers:

$$Votes = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper$$

Results:

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	-0.001	0.0059	-0.18	0.8599

Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistic	570

The residual standard error **RSE** is the standard deviation of the errors:

$$RSE = \sqrt{MS_{residuals}} = \sqrt{\frac{RSS}{df_{residual}}} = \sqrt{\frac{\sum_{i=0}^{N}(\hat{y}_i - y_i)^2}{N-p-1}}$$

Categorical Predictors

Until now, we have only considered continuous predictors. We can also use categories as predictors:

We use **dummy variables**: Represent categories with only zeros and ones.

Idea:
Select one category as the baseline, for the rest of the categories create a dummy variable. Each category will be represented with a one in their dummy variable and a zero in the others. For the baseline category all zeros.

Categorical Predictors

Example: We want to predict the Pokemon attack with the help of Pokemon types. The categories are: normal, water, fire and fairy.

	Dummy Variable 1	*Dummy Variable 2*	*Dummy Variable 3*
Water	1	0	0
Fire	0	1	0
Fairy	0	0	1
Normal	0	0	0

$$Attack = \beta_0 + \beta_1 Dummy_1 + \beta_2 Dummy_2 + \beta_3 Dummy_3$$

Categorical Predictors

After creating the linear model, the F-test will say if the model is significant.

The t-tests will tell if there is a difference between a category and the baseline category you selected. The t-test compare... the mean between groups!

Example continued: The meaning of the coefficents for the linear model are...

$$Attack = \beta_0 + \beta_1 Dummy_1 + \beta_2 Dummy_2 + \beta_3 Dummy_3$$

- β_0 is the mean of the attack of Pokemons type Normal.
- β_1 is the difference between the attack of Pokemons of type Normal and Water.
- β_2 is the difference between the attack of Pokemons of type Normal and Fire.
- β_3 is the difference between the attack of Pokemons of type Normal and Fairy.

Categorical Predictors

Is it now more clear why the statistical tests for regression are the same as for one-way ANOVA + posthoc tests!

All the statistical tests to compare means that we learned on week 4 are actually special cases of regression!

Ordinal Predictors

In case that we have ordinal categories, it is better to transform to numerical values and no dummy variables.

Example: We have Airbnb accomodations: One room, complete apartment, luxury apartment, house, mansion.

We can give numbers to each category: One room (1), complete apartment (5), luxury apartment (10), house (15), mansion (20).

Assumptions

Same assumptions as with linear regression All of them.

One more important assumption with several predictors:

* No perfect **multicollinearity**: There should be no perfect linear relationship between two or more of the predictors. So, the predictor variables should not correlate too highly.

You want predictor variables that are related to the outcome variable but unrelated to each other.

Multicollinearity

Multicollinearity exists when there is a strong correlation between two or more predictors in a regression model. It creates untrustworthy βs and reduces the size of R.

Example: We want to predict the height of people using as predictors the size of the right leg and left leg

$$Height = \beta_0 + \beta_1 RightLeg + \beta_2 LeftLeg$$

The coefficients will most likely be not significant and the R^2 will be low...

To test for multicollinearity, we calculate the VIF measure for all predictors:

- Predictors with VIF values larger than 10 are a serious problem.
- If the average VIF for all predictors is substantially greater than 1 then the regression may be biased.

Selecting the Best Model

How to pick the best predictors? = How to choose the best model?

We want to build a linear model that makes the best predictions. Similar idea as when we wanted to pick the best distribution to fit our data on Week 7.

Idea 1: Choose the model that has the highest R^2 or lowest RSS.

However, we can only use R^2 and RSS to compare models with the same number of predictors, since they always decrease the more predictors we include in the model.

Selecting the Best Model

How to pick the best predictors? = How to choose the best model?

We want to build a linear model that makes the best predictions. Similar idea as when we wanted to pick the best distribution to fit our data on Week 7.

Idea 1: Choose the model that has the highest R^2 or lowest RSS.

However, we can only use R^2 and RSS to compare models with the same number of predictors, since they always decrease the more predictors we include in the model.

Idea 2: We use **AIC or BIC** to compare between models of different number of predictors! The same as when comparing distributions, because they take into account the number of predictors.

$$AIC = N * log\frac{RSS}{N} + 2 * P \qquad BIC = N * log\frac{RSS}{N} + log(N) * P$$

Selecting the Best Model

We can compare all possible combinations and find the best model. However for many predictors the combinations increase exponentially.

Example We want to predict the level of headache I will have depending on the number of vodka,tequila and whisky shots I took last night.

$$Headache = \beta_0$$

$$Headache = \beta_0 + \beta_1 Vodka$$

$$Headache = \beta_0 + \beta_2 Tequila$$

$$Headache = \beta_0 + \beta_3 Whysky$$

$$Headache = \beta_0 + \beta_1 Vodka + \beta_2 Tequila$$

$$Headache = \beta_0 + \beta_2 Tequila + \beta_3 Whysky$$

$$Headache = \beta_0 + \beta_1 Vodka + \beta_2 Whisky$$

$$Headache = \beta_0 + \beta_1 Vodka + \beta_2 Tequila + \beta_3 Whysky$$

Example For 10 predictors, you have 1024 combinations...

Selecting the Best Model

What to do with many predictors? Stepwise selection methods:

Forward Stepwise Selection:

1. Start with the basic model with only β_0.
2. Add the predictor that reduces the RSS the most.
3. Repeat step 2 until you have all the predictors.
4. At the end you have one model with one predictor, one with two predictors, one with three and so on. Select the one with the lowest AIC as the best model.

Selecting the Best Model

What to do with many predictors? Stepwise selection methods:

Backward Stepwise Selection:

1. Start with the complete model with all the predictors.
2. Remove the predictor that increases the RSS the **least**.
3. Repeat step 2 until you have the basic model with only β_0.
4. At the end you have one model with one predictor, one with two predictors, one with three and so on. Select the one with the lowest AIC as the best model.

Selecting the Best Model

It is also possible to combine forward steps with backward steps, to get an even better model.

Important Note For two models with similar AIC, it is always preferable to select the model with less predictors. Models with many predictors increase the variability of the predictions and induce overfitting.

More on this next week...

Literature

- Discovering Statistics using R: Chapter 7