Presentation Scripts

(Jun) intro and data description -------

hi i am ——. today my teammate and me are going to help you understand the house price in science.
&&&
We plan to go over the Introduction and Data description
Analyze the Appropriate model selection with Four Assumptions
Conclusions and Limitations
&&
Now let's talk about the introduction and Data description &&
the goal of our presentation is to find a way to balance people's wealth and numerous elements that can affect the house prices.)
and predict the house price with the dataset by multiple regression
&&
And we got a data set about Boston house price with 506 cases and 14 attributes from the U.S Census Service
also, one of the photomask for this dataset is to predict the house price.(48)
I know you might be confused to the table, now is the metadata for the dataset &&
to other 13 includes the criminal rate, the proportion of residential land zone, non-retail business acres, whether the house tract bounds the Charles River, nitric oxides concentration, number of rooms ,distances to centers, accessibility to radial highways, property-tax rate, pupil-teacher ratio, the proportion of blacks, percentage of lower status population and finally, the line in red, the median house price that we want to set as the dependent variable as MEDV. (1.24)

(Yifei) assumption checking and log transformation ----

&& hi, i am Yoffie. As for the assumption checking part,
We first created the && multiple regression of Boston house price on the other 13 variables and tested the assumption. We used autoplot() to draw residuals versus fitted plots and normal qq plot. As first plot shows the non-linear pattern because the residuals are above zero for low house price, then they go below zero and end up again above zero for high house price just like a smiley face,
So, we used the log transformation. (1.54) &&
After log transformation for MEDV, we now get a Residuals vs fitted plot without obvious pattern. Therefore, we were able to satisfy the linearity assumption.
&&Moreover, we met the heteroskedasticity assumption too. As each observation is independent of the other observations, we satisfy the independence assumption. Even though the residuals appear to be fanning out on the region that fitted values are greater than 3.0, the general spread looks reasonably constant over the fitted values.
The qq plot shows the non-linear trend, but as we have 506 large enough samples, we rely on Central Limit Theorem to satisfy the normality assumption. All assumptions for log transformation passed, we can use log-transformed data.(2.51

(seunghyun) model creation and hypothesis testing ----------

hi I'm seunghyun (backward search using AIC) Having full model as a base line, we used backward search/ using AIC with step() function. This method dropped the INDUS and AGE variables and leave 11 explanatory variables with adjusted r-squared 0.78. &&
(forward search using AIC) Having model with no explanatory variables as a base line, we also used forward search/ using AIC. The result included same 11 explanatory variables with adjusted r-squared 0.78.
(final AIC model formula) As backward stepwise AIC model and forward stepwise AIC model contains exactly same 11 explanatory variables , we believe that our model is stable. (3.43)
<hypothesis testing> &&
(1st hypothesis testing) For the AIC- related model, we conducted three hypothesis testing to see if dropping the AGE and INDUS was statistically reasonable choice. For the first hypothesis testing, we wanted to test whether the coefficient for AGE is significantly different to zero in the multiple regression model.
Letting the full model as equation shown in the slide, we set our null hypothesis as Beta 13 is equal to 0 and alternative hypothesis as $\beta 13 != 0$. The assumptions were met for the full model. We had 0.691 as p-value, which is greater than 0.05, and therefore we do not reject null hypothesis at the 5% level of significance. Hence, there is no evidence to suggest that there is a significant linear relationship between MEDV, and AGE and it can be dropped from the model &&
We took same process with reduced 12 attribute model to test whether we should drop INDUS which was a variable that has next biggest p-value. From this hypothesis testing, we do not reject null hypothesis and therefore dropped INDUS from the model.&& Another hypothesis test was taken with model that has 11 attribute to test whether we should drop ZN. In this test, we reject the null hypothesis , and therefore ZN cannot be dropped. Through these three-hypothesis testing, we can conclude that dropping AGE and INDUS from AIC model was reasonable choice. (4.45)

(Baiheng) model selection and key conclusion---------- ----------

&& hi there is bill
To evaluate which model is best between our step model with 11 variable, full model with 13 variable, and simple model with only RM as its variable as an example
we used caret package and did cross validation with 10 repetitions to further investigate. We calculated the RMSE, MAE, and R-squared for each trial and for each model
. according to the plot, Since the created model has slightly smaller RMSE and MAE than full model, and also slightly larger R-squared, we chose the step model
and this is the model and formula
we also need to do assumptions to this model to make sure it can work
this plot looks similar to the full model's one, so assumptions for this model are all met, too (5.33)
now we are in part 3
let's interpret the formula of our model. Our model is a Log-linear mode
so, one unit increase in x will result in a coefficient beta * 100% change in Y.

actually, there are 11 conclusions because we have 11 different variables
let’s just check the first two with biggest influence on house price
nitric oxides concentration results in a 72-percentage decrease in house price on average
and if house nearby the river can result in 10% percentage in house price on average (6.10)

(key conclusion)
For the key conclusion,
 the factors that *decreases* the price are these 5 like crime rate
and the factors that *increases* the price are these 6 like room number
among them,
NOX and CHAS are variables that affect the house price the most,
we can infer that the demand for the clean air and nature view was high at the moment that this data was collected,(6.40)

(zijun) limitation ----------

(limitation)
This analysis has 3 limitations. First one is about Assumption for Homoscedasticity. Although the residual and fitting plot shows a fairly random plot. So further tests are needed.
Second one is about Bounded MEDV. 16 cases that have 50.00 at their MEDV might have a higher price than $50000. Therefore, further research with precise house price is needed to get the model that is better representing the Boston house price. (7.30)
Last one is related to a small sized data set. The data set has 506 cases, which is large enough to make the model statistically reliable and representative. However, as the data sheet points out, it is not enough to represent the housing prices in Boston as a whole. (8.0)
Although we use 1978 data that is quite outdated, there are some limitations because the current high housing prices hurt many people. We hope that our analysis will help find a balance between people’s wealth and many factors that can affect house price. Thank you! (8.20)
