Data Analysis 2021 Spring





# Lecture 06:

**Linear Regression II**

##### April 7 & April 12, 2021

**Taesoo Kwon**

[tskwon@seoultech.ac.kr](mailto:tskwon@seoultech.ac.kr)

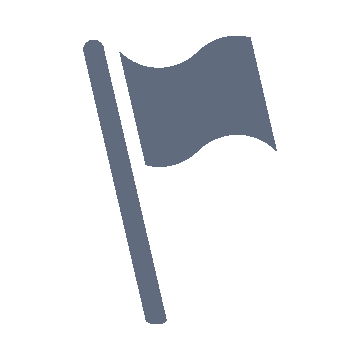
©copyright CIEL 2015 https://ciel.seoultech.ac.kr/

**Course Schedule (Tentative)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Week** | **Topics** | **Note** | **Date (W)** | **Date (M)** |
| 1 | Orientation, Statistical Learning (Ch2) | Online | 03/03 | 03/08 |
| 2 | Statistical Learning (Ch2), Python Programming | Online | 03/10 | 03/15 |
| 3 | Probability & Statistics | Online | 03/17 | 03/22 |
| 4 | Probability & Statistics | Online | 03/24 | 03/29 |
| 5 | Linear Regression (Ch3) | Online | 03/31 | 04/05 |
| **6** | Linear Regression (Ch3) | Online | 04/07 | 04/12 |
| 7 | Classification (Ch4) | Online | 04/14 | 04/19 |
| 8 | **Midterm exam** | **7pm or Class hours (W1-W7)** | **04/21or26** | **04/21or26** |
| 9 | Resampling Methods (Ch5) | Online | 04/28 | 05/03 |
| 10 | Linear Model Selection and Regularization (Ch6) | Online | 05/05 | 05/10 |
| 11 | Moving Beyond Linearity (Ch7) | Online | 05/12 | 05/17 |
| 12 | Tree-Based Methods (Ch8) | Online | 05/19 | 05/24 |
| 13 | Support Vector Machines (Ch9) | Online | 05/26 | 05/31 |
| 14 | Unsupervised Learning (Ch10) | Online | 06/02 | 06/07 |
| 15 | **Final exam** | **7pm or Class hours (W9-W14)** | **06/09or14** | **06/09or14** |

©copyright CIEL 2015 **2**

#### Multiple linear regression



**OUTLINES**

* Other considerations in regression model
* Python lab
* Summary & Next class

©copyright CIEL 2015



# Multiple Linear Regression



**: Ch3.2**

### Multiple linear regression

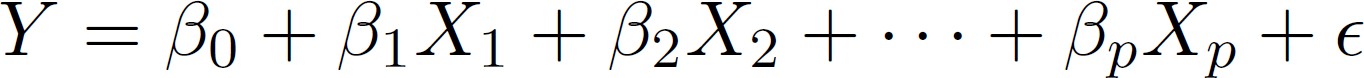
#### Other considerations in regression model

* + Python lab
  + Summary & Next class

©copyright CIEL 2015 https://ciel.seoultech.ac.kr/

**Multiple Linear Regression**

#### Here our model is



* We interpret 𝛽𝛽𝑗𝑗 as the average effect on 𝑌𝑌 of a one unit increase in 𝑋𝑋𝑗𝑗 , holding all other predictors fixed.
  + In the advertising example, the model becomes



©copyright CIEL 2015 **5**

## Interpreting Regression Coefficients

#### The ideal scenario is when the predictors are uncorrelated  a balanced design

* + Each coefficient can be estimated and tested
  + Interpretations such as “a unit change in 𝑋𝑋𝑗𝑗 is associated with a 𝛽𝛽𝑗𝑗 change in 𝑌𝑌, while all the other variables stay fixed", are possible.

#### Correlations amongst predictors cause problems

* + The variance of all coefficients tends to increase, sometimes dramatically
  + Interpretations become hazardous
    - When 𝑋𝑋𝑗𝑗 changes, everything else changes

#### Claims of causality should be avoided for observational data

* E.g., a regression of shark attacks versus ice cream sales

©copyright CIEL 2015 **6**

## Estimation and Prediction for Multiple Regression

1

𝑝𝑝

#### Given estimates

0

𝛽𝛽̂

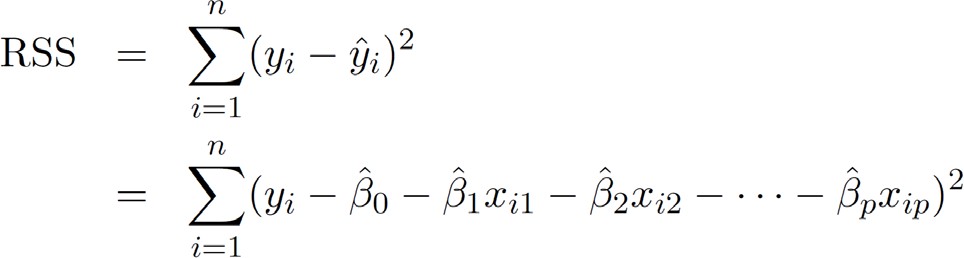
, 𝛽𝛽̂

, ⋯ 𝛽𝛽̂

, we can make predictions using the formula



* We estimate 𝛽𝛽0, 𝛽𝛽1, ⋯ , 𝛽𝛽𝑝𝑝 as the values that minimize the sum of squared residuals



* + This is done using standard statistical software
  + The values

0

1

𝛽𝛽̂

, 𝛽𝛽̂

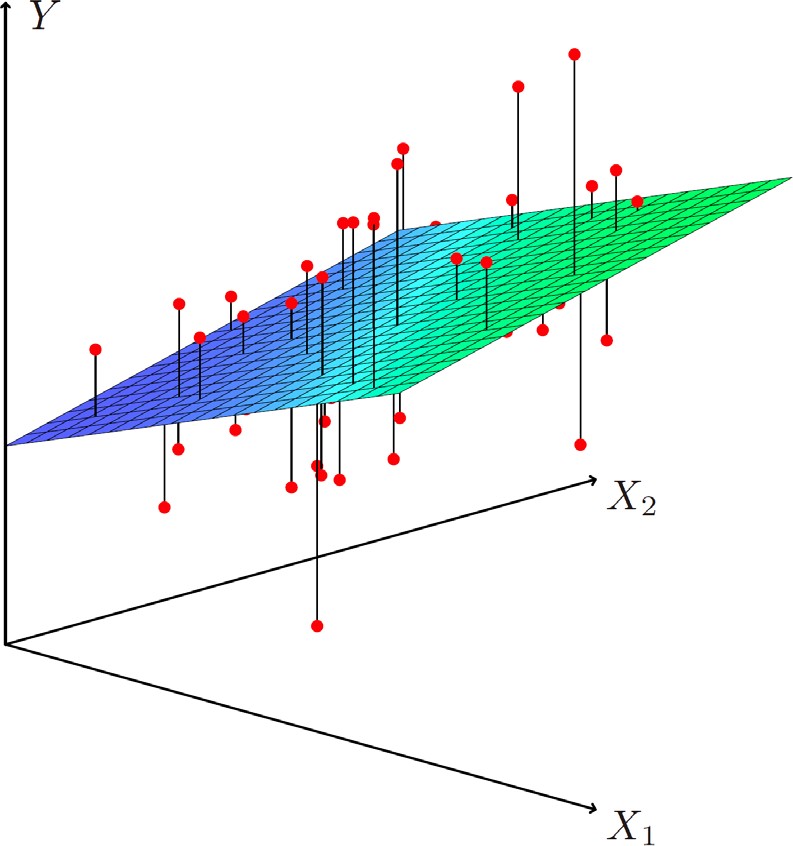
, ⋯ 𝛽𝛽̂

that minimize RSS are the multiple least squares regression coefficient estimates.

©copyright CIEL 2015 **7**

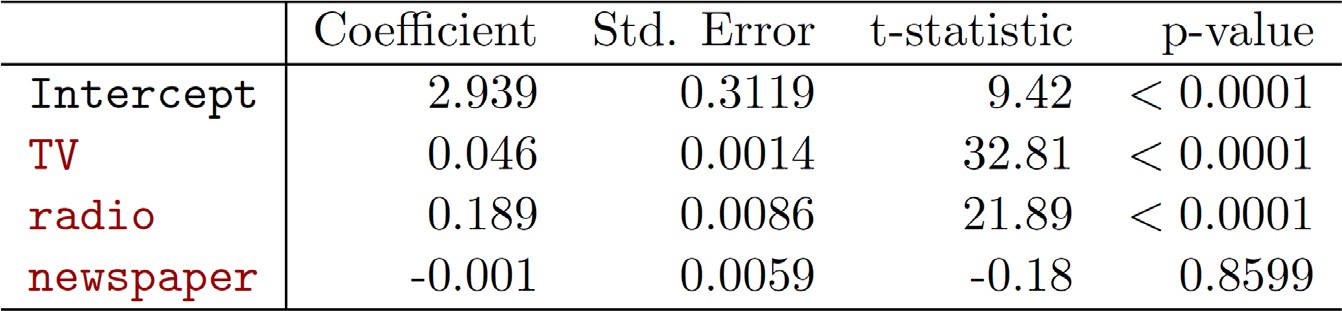
𝑝𝑝

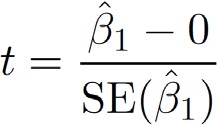
## Example of Two Predictors and One Response



©copyright CIEL 2015 **8**

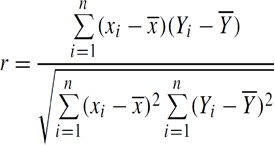
## Results for Advertising Data







Ross



©copyright CIEL 2015 **9**

## Cf. Simple Linear Regression vs. Multiple Linear Regression

#### Simple linear regression

* + Ignoring other predictors
  + newspaper: meaningful?



©copyright CIEL 2015 **10**

## Some Important Questions

#### Is at least one of the predictors 𝑋𝑋1, 𝑋𝑋2, ⋯ , 𝑋𝑋𝑝𝑝 useful in predicting the response?

1. Do all the predictors help to explain 𝑌𝑌 , or is only a subset of the predictors useful?
2. How well does the model fit the data?
3. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

©copyright CIEL 2015 **11**

## Q1: Is At Least One Predictor Useful?

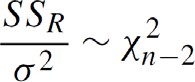
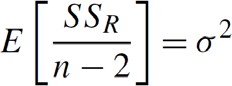
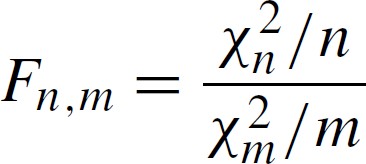
#### Relationship between response and predictors?  Using hypothesis testing

* + Null hypothesis



* + - Alternative hypothesis
  + Tested by computing 𝐹𝐹-statistics

Ross



𝐻𝐻0 true:



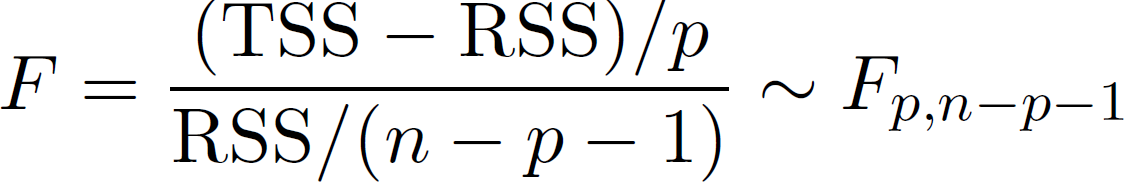
𝐹𝐹𝑝𝑝,𝑛𝑛−𝑝𝑝−1 = 1 ?

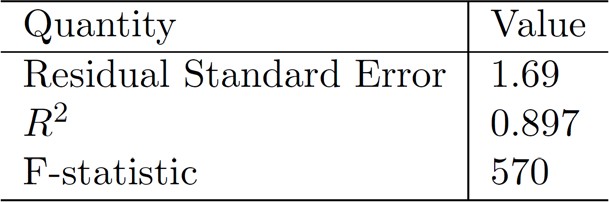
𝐻𝐻𝑎𝑎 true: 

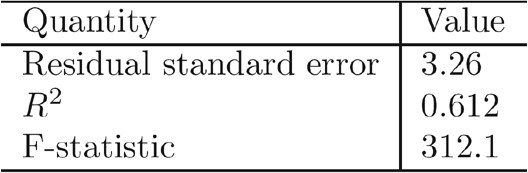
©copyright CIEL 2015 **12**

## Q1: Is At Least One Predictor Useful? [cont.]

#### For the first question, we can use the 𝐹𝐹-statistic



* + Regression of number of units sold on TV, newspaper, and radio
    - Regression of number of units sold on only TV



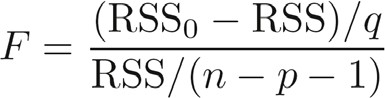
©copyright CIEL 2015 **13**

## Q1: Is At Least One Predictor Useful? [cont.]

#### Testing that a particular subset of 𝑞𝑞 of coefficients are zero



* + RSS0: residual sum of squares for the model that uses all the variables except those last 𝑞𝑞
  + 𝐹𝐹-statistic



* + For omitting a single variable (i.e., 𝑞𝑞 = 1): 𝑡𝑡-statistic
  + However, for a large 𝑝𝑝, 𝐹𝐹-statistic required eventually

#### What if 𝑝𝑝 > 𝑛𝑛?  high-dimensional setting

* + 𝐹𝐹-statistic cannot be used)
  + Solution: e.g., forward selection (Ch 6)

©copyright CIEL 2015 **14**

## Q2: Deciding on Important Variables

#### The most direct approach is called all subsets or best subsets regression

* + We compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size

#### However we often can't examine all possible models, since they are 2𝑝𝑝 of them

* + E.g., when 𝑝𝑝 = 40, there are over a billion models!

#### Instead we need an automated approach that searches through a subset of them

* + We discuss two commonly use approaches next

©copyright CIEL 2015 **15**

## Q2: Deciding on Important Variables [cont.]

#### Forward selection

* + Begin with the null model
    - A model that contains an intercept but no predictors
  + Fit 𝑝𝑝 simple linear regressions and add to the null model the variable that results in the lowest RSS
  + Add to that model the variable that results in the lowest RSS amongst all two-variable models.
  + Continue until some stopping rule is satisfied
    - E.g., when all remaining variables have a 𝑝𝑝-value above some threshold.
  + Forward selection can always be used. It is a greedy approach
    - It might include variables early that later become redundant

©copyright CIEL 2015 **16**

## Q2: Deciding on Important Variables [cont.]

#### Backward selection

* + Start with all variables in the model
  + Remove the variable with the largest 𝑝𝑝-value
    - That is, the variable that is the least statistically significant.
  + The new -variable model is fit, and the variable with the largest 𝑝𝑝-value is removed

𝑝𝑝 − 1

* + Continue until a stopping rule is reached.
    - For instance, we may stop when all remaining variables have a significant 𝑝𝑝-value defined by some significance threshold

#### Later we discuss more systematic criteria for choosing an “optimal" member in the path of models produced by forward or backward stepwise selection

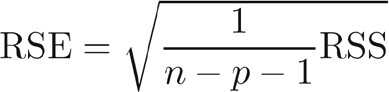
©copyright CIEL 2015 **17**

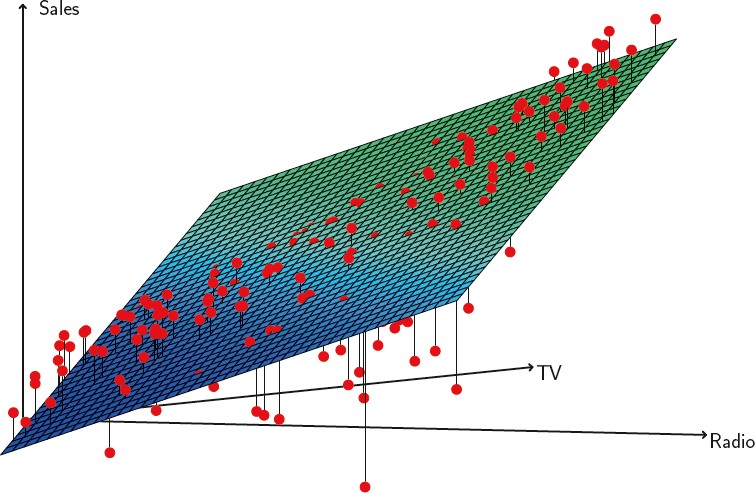
## Q3: Model Fit: How Well does Model Fit data?

##### 𝑅𝑅 squared (𝑅𝑅2)

* + 𝑅𝑅2 always increase when more variable are added to the model, even if those variables are only weakly associated with the response
    - E.g., adding newspaper  a tiny increase in 𝑅𝑅2

##### RSE



* + Models with more variables can have higher RSE if the decrease in RSS is small relative to the increase in 𝑝𝑝
* Graphical summaries by plotting data
  + Non-linear pattern cannot be modeled accurately using linear regression
  + Synergy and interaction effect
    - E.g., between advertising media whereby combining the media together results in a bigger boost to sales than using any single medium

©copyright CIEL 2015 **18**

## Q4: Predictions: What Response Value and How Accurate

#### Confidence interval

* + Inference about 

#### Prediction interval

* + Inference about 
  + Prediction intervals are always wider than confidence intervals

©copyright CIEL 2015 **19**



# Other Considerations in Regression Model: Ch3.3



#### Multiple linear regression

* Other considerations in regression model
* Python lab
* Summary & Next class

©copyright CIEL 2015 https://ciel.seoultech.ac.kr/

**Other Consideration in Regression Model**

#### Qualitative predictors

* + Dummy variables

#### Extension of linear model

* + Synergy or interaction effect

©copyright CIEL 2015 **21**

## Qualitative Predictors

#### Some predictors are not quantitative but are qualitative, taking a discrete set of values

* These are also called categorical predictors or factor variables
* See for example the scatterplot matrix of the credit card data in the next slide
  + In addition to the 7 quantitative variables shown, there are four qualitative variables: gender, student

(student status), status (marital status), and ethnicity (Caucasian, African American (AA) or Asian)

©copyright CIEL 2015 **22**

## Qualitative Predictors [cont.]

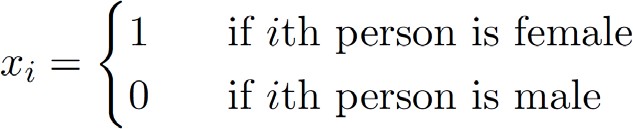
#### Credit card data

©copyright CIEL 2015 **23**

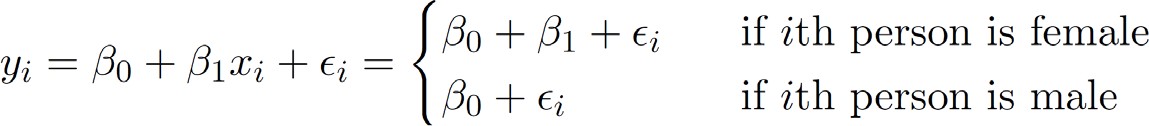
## Qualitative Predictors [cont.]

#### Example: investigate differences in credit card balance between males and females, ignoring the other variables

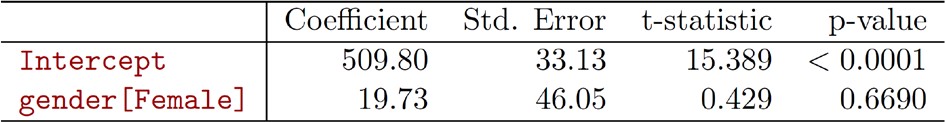
* + We create a new variable



* + Resulting model



* + Results for gender model

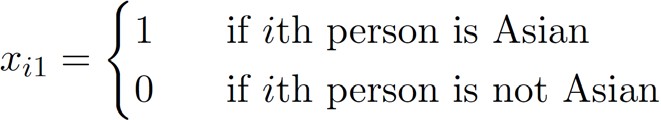


©copyright CIEL 2015 **24**

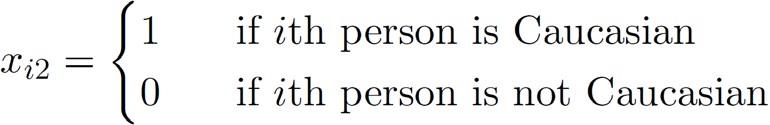
## Qualitative Predictors [cont.]

#### Qualitative predictors with more than two levels

* + With more than two levels, we create additional dummy variables
    - For example, for the ethnicity variable we create two dummy variables
    - The first could be



* + - And the second could be

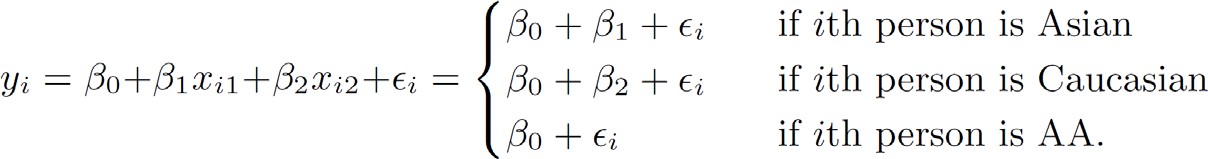


©copyright CIEL 2015 **25**

## Qualitative Predictors [cont.]

#### Qualitative predictors with more than two levels

* + Then both of these variables can be used in the regression equation, in order to obtain the model



* + There will always be one fewer dummy variable than the number of levels
    - The level with no dummy variable, i.e., African American in this example, is known as the baseline.



©copyright CIEL 2015 **26**

## Extensions of Linear Model

#### Removing the additive assumption: interactions and nonlinearity

* Interactions
  + In our previous analysis of the Advertising data, we assumed that the effect on sales of increasing one advertising medium is independent of the amount spent on the other media
  + For example, the linear model



* + - It states that the average effect on sales of a one-unit increase in TV is always 𝛽𝛽1, regardless of the amount spent on radio

©copyright CIEL 2015 **27**

## Extensions of Linear Model [cont.]

#### Interactions [cont.]

* + But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases
  + In this situation, given a fixed budget of $100,000, spending half on radio and half on TV may increase

sales more than allocating the entire amount to either TV or to radio

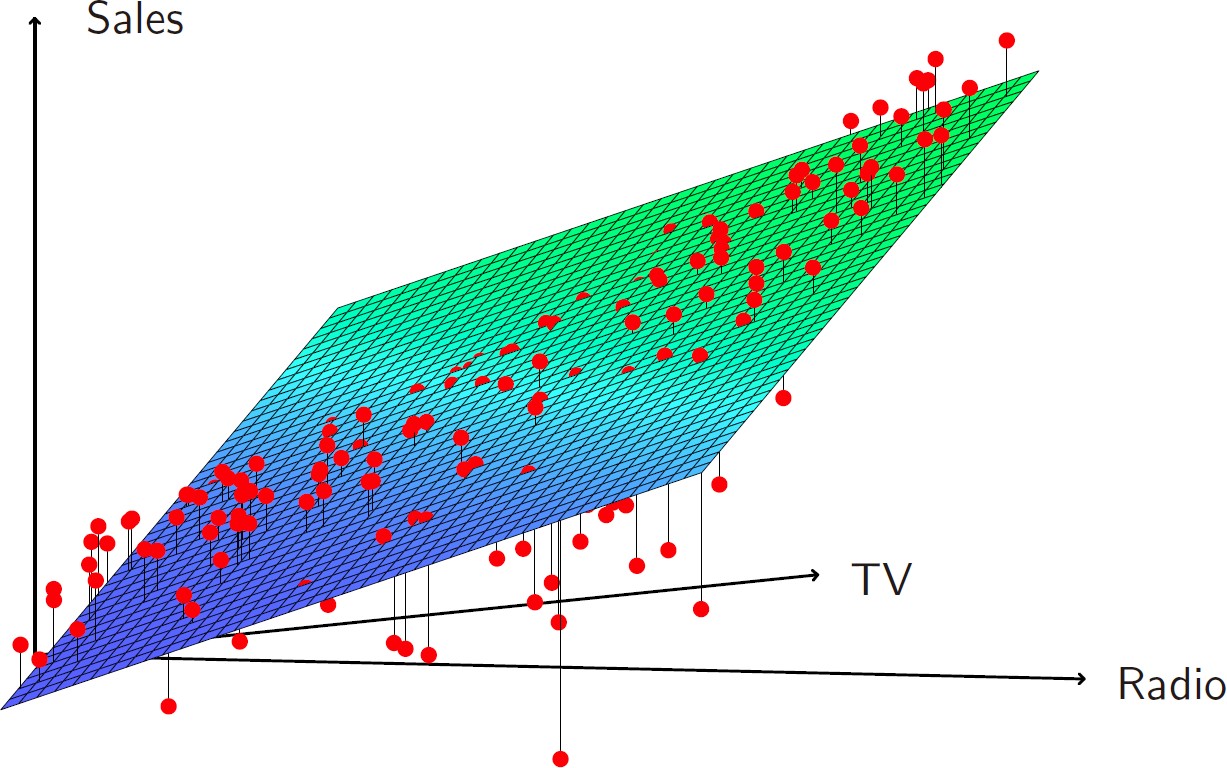
* + In marketing, this is known as a synergy effect, and in statistics it is referred to as an interaction effect

©copyright CIEL 2015 **28**

## Extensions of Linear Model [cont.]

#### Interaction in the Advertising data?

* + When levels of either TV or radio are low, then the true sales are lower than predicted by the linear model
  + But when advertising is split between the two media, then the model tends to underestimate sales

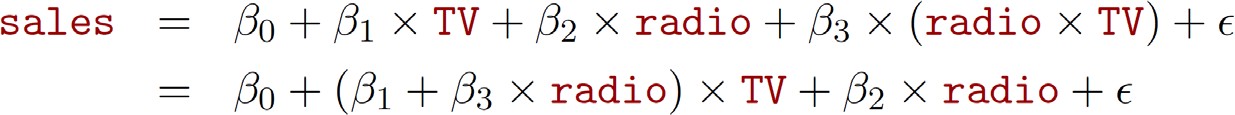


©copyright CIEL 2015 **29**

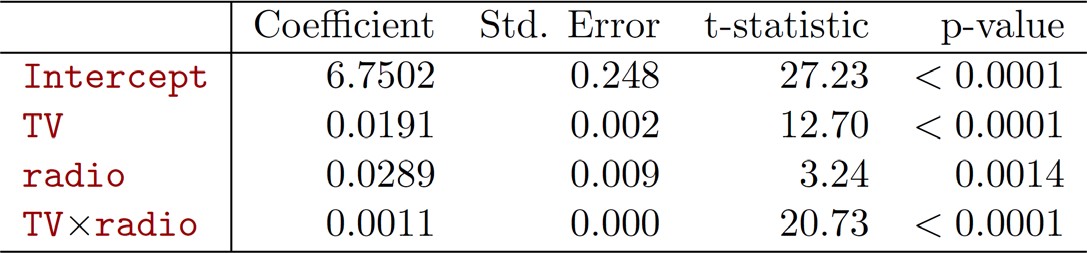
## Extensions of Linear Model [cont.]

#### Modelling interactions: Advertising data

* + Model takes the form



* + Results



©copyright CIEL 2015 **30**

## Extensions of Linear Model [cont.]

#### Interpretation

* + The results in this table suggests that interactions are important
  + The 𝑝𝑝-value for the interaction term TV×radio is extremely low, indicating that there is strong evidence for 𝐻𝐻𝐴𝐴: 𝛽𝛽3 ≠ 0
  + The 𝑅𝑅2 for the interaction model is 96.8%, compared to only 89.7% for the model that predicts sales

using TV and radio without an interaction term

* + This means that (96.8 – 89.7)/(100 – 89.7) = 69% of the variability in sales that remains after fitting the additive model has been explained by the interaction term
  + The coefficient estimates in the table suggest that an increase in TV advertising of $1,000 is associated with increased sales of



* + An increase in radio advertising of $1,000 will be associated with an increase in sales of



©copyright CIEL 2015 **31**

## Extensions of Linear Model [cont.]

#### Hierarchy

* + Sometimes it is the case that an interaction term has a very small 𝑝𝑝-value, but the associated main effects (in this case, TV and radio) do not
  + Hierarchy principle
    - If we include an interaction in a model, we should also include the main effects, even if the 𝑝𝑝-values associated with their coefficients are not significant
  + The rationale for this principle is that interactions are hard to interpret in a model without main effects
    - Their meaning is changed
  + Specifically, the interaction terms also contain main effects, if the model has no main effect terms.

©copyright CIEL 2015 **32**

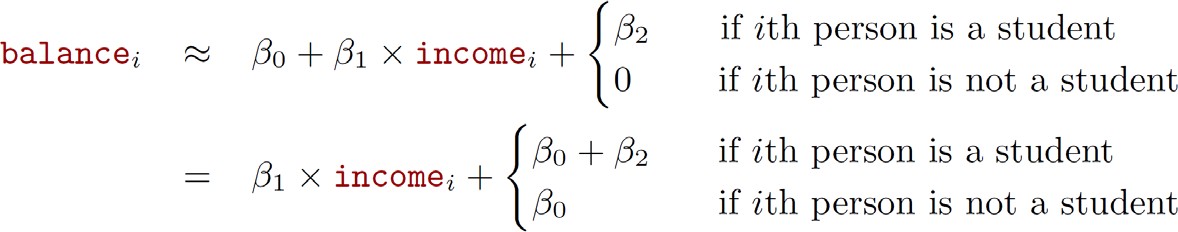
## Extensions of Linear Model [cont.]

#### Interactions between qualitative and quantitative variables

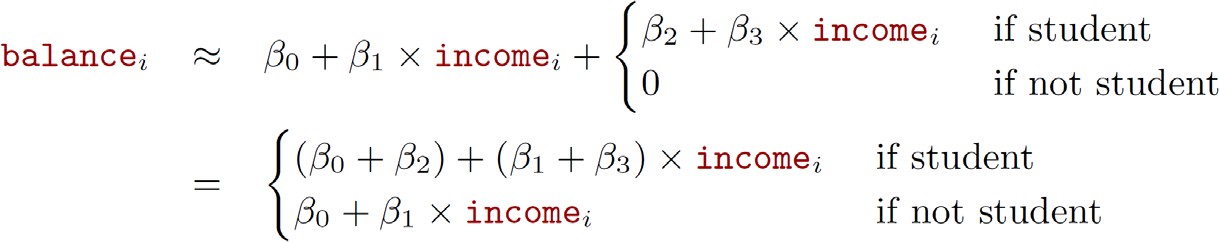
* + Consider the Credit data set, and suppose that we wish to predict balance using income (quantitative) and

student (qualitative)

* + Without an interaction term, the model takes the form



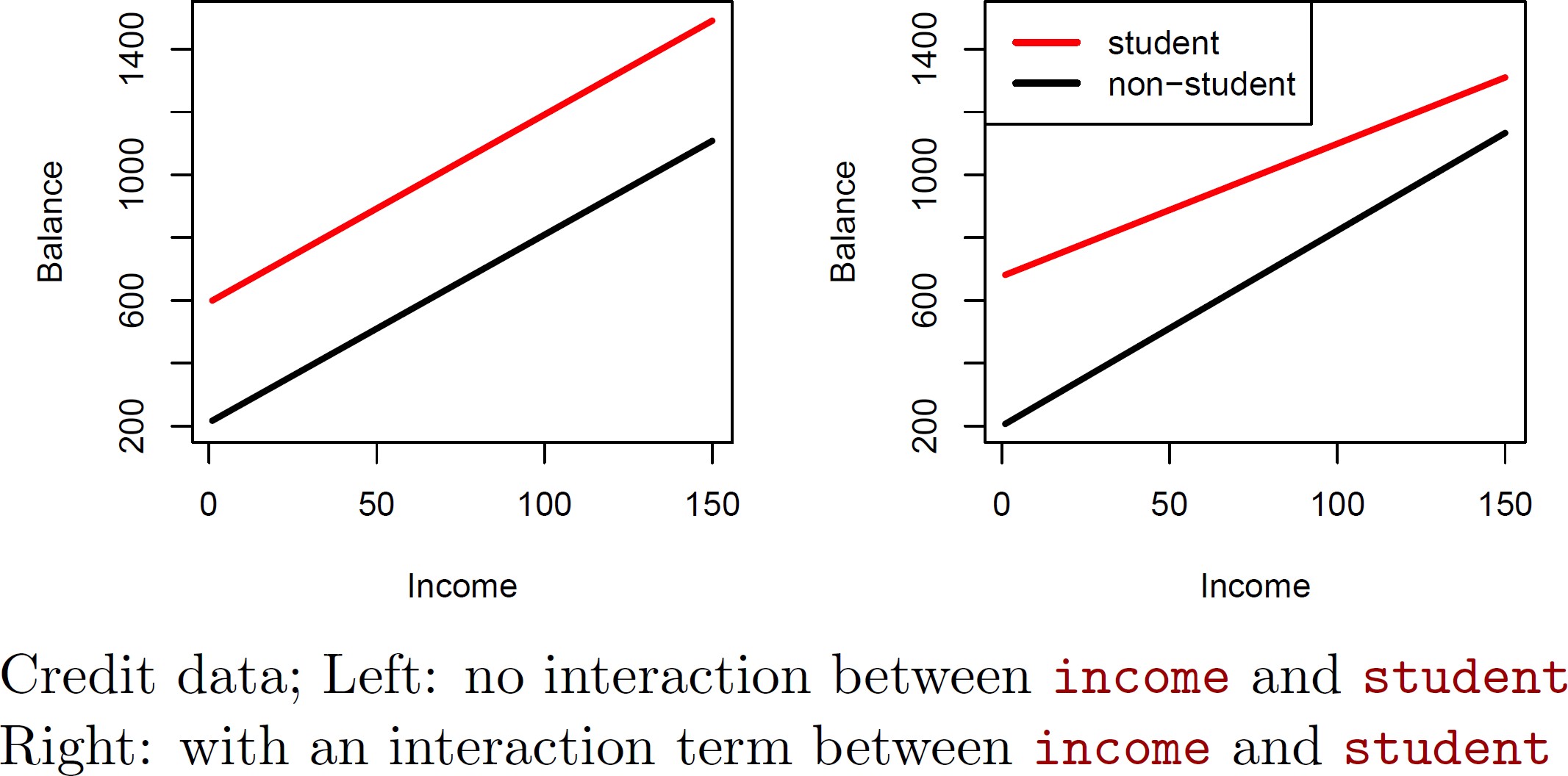
* + With interactions, it takes the form



©copyright CIEL 2015 **33**

## Extensions of Linear Model [cont.]

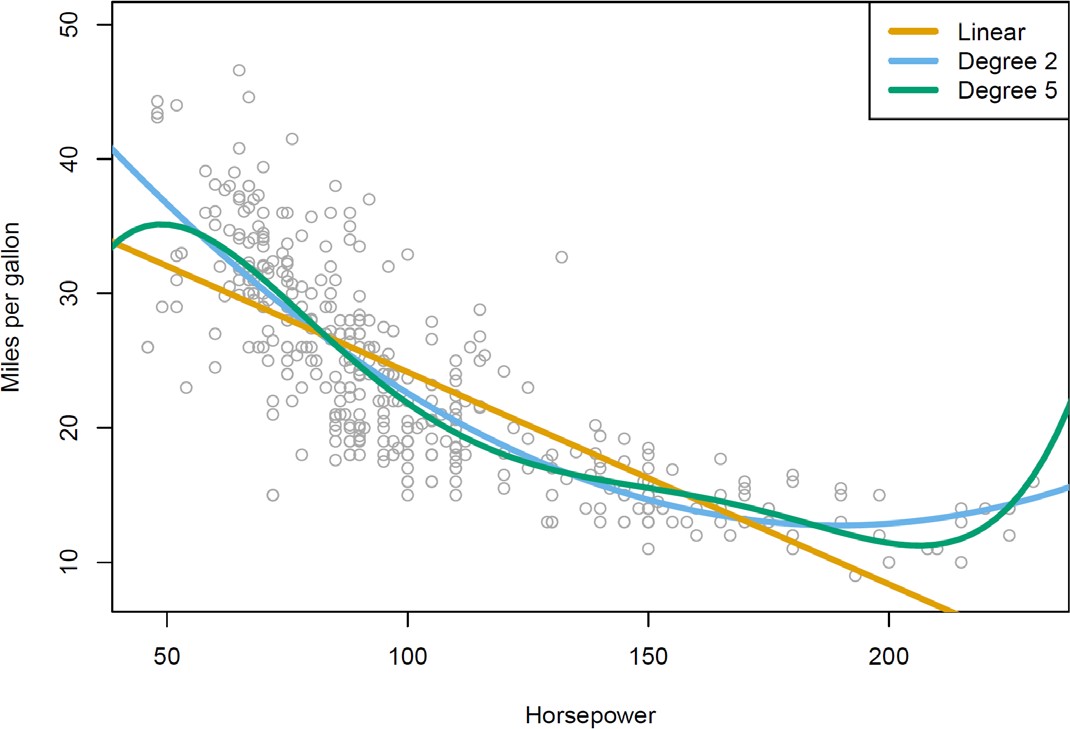
#### Interactions between qualitative and quantitative variables [cont.]



©copyright CIEL 2015 **34**

## Non-linear Effects of Predictors

#### Polynomial regression on Auto data



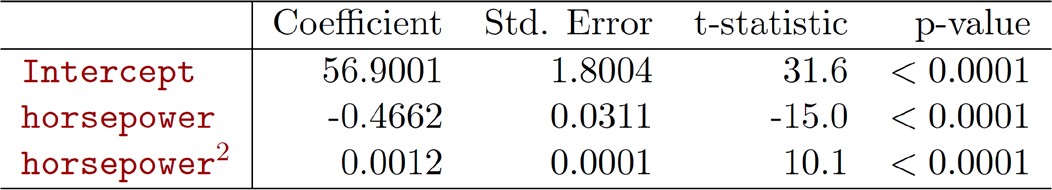
©copyright CIEL 2015 **35**

## Non-linear Effects of Predictors

#### The figure suggests that



* + It may provide a better fit



©copyright CIEL 2015 **36**

## Potential Problems

#### Non-constant variance of error terms

* Outliers
* High leverage points
* Collinearity

©copyright CIEL 2015 **37**



**Python Programming**

#### Multiple linear regression

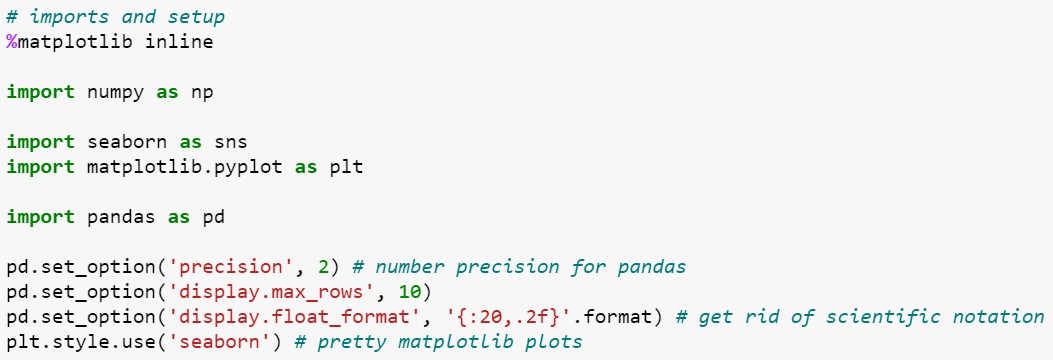
* Other considerations in regression model
* Python lab
* Summary & Next class

©copyright CIEL 2015 https://ciel.seoultech.ac.kr/

## Import

#### Using Python Libraries

* + - Import the libraries that are often used for data analysis



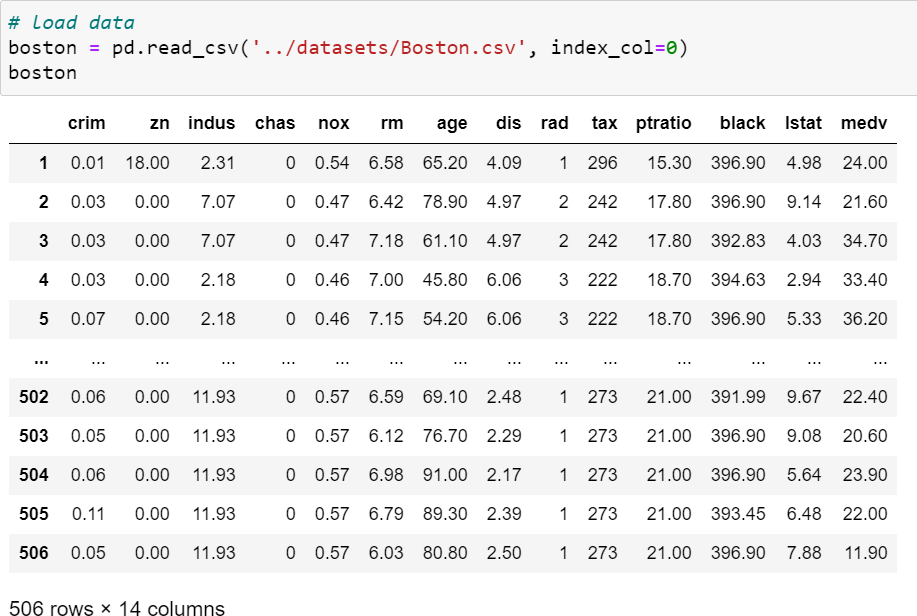
Display float values to two decimal places

©copyright CIEL 2015 **39**

## Load Datasets

Average number of rooms per house

Average age of houses



median house value

506 neighborhoods around Boston

percent of households with low socioeconomics status

©copyright CIEL 2015 **40**

## Scikit-learn Libraries

* + scikit-learn: <https://scikit-learn.org/>
    - Simple and efficient tools for predictive data analysis
    - Built on NumPy, SciPy, and matplotlib

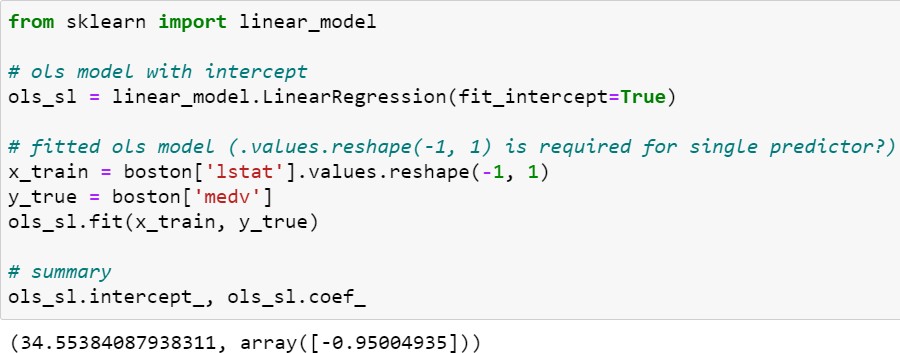
##### Classsification: Identifying which category an object belongs to

* + Regression: Predicting a continuous-valued attribute associated with an object
  + Clustering: Automatic grouping of similar objects into sets
  + Dimensionality reduction: Reducing the number of random variables to consider
  + Model selection: Comparing, validating and choosing parameters and models
  + Preprocessing: Feature extraction and normalization

©copyright CIEL 2015 **41**

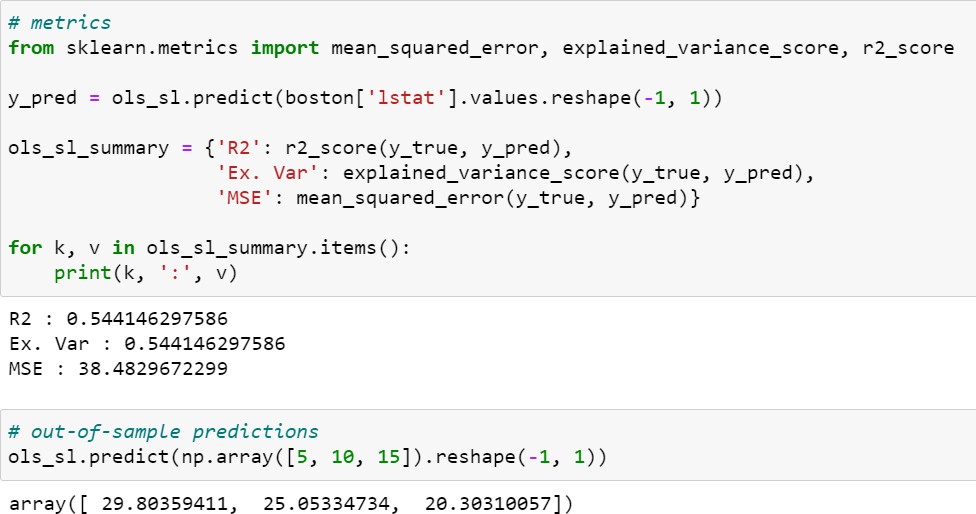
## 3.6.2 Simple Linear Regression

#### Using scikit-learn



Whether to calculate the intercept for this model

the value is inferred from the length of the array and remaining dimensions

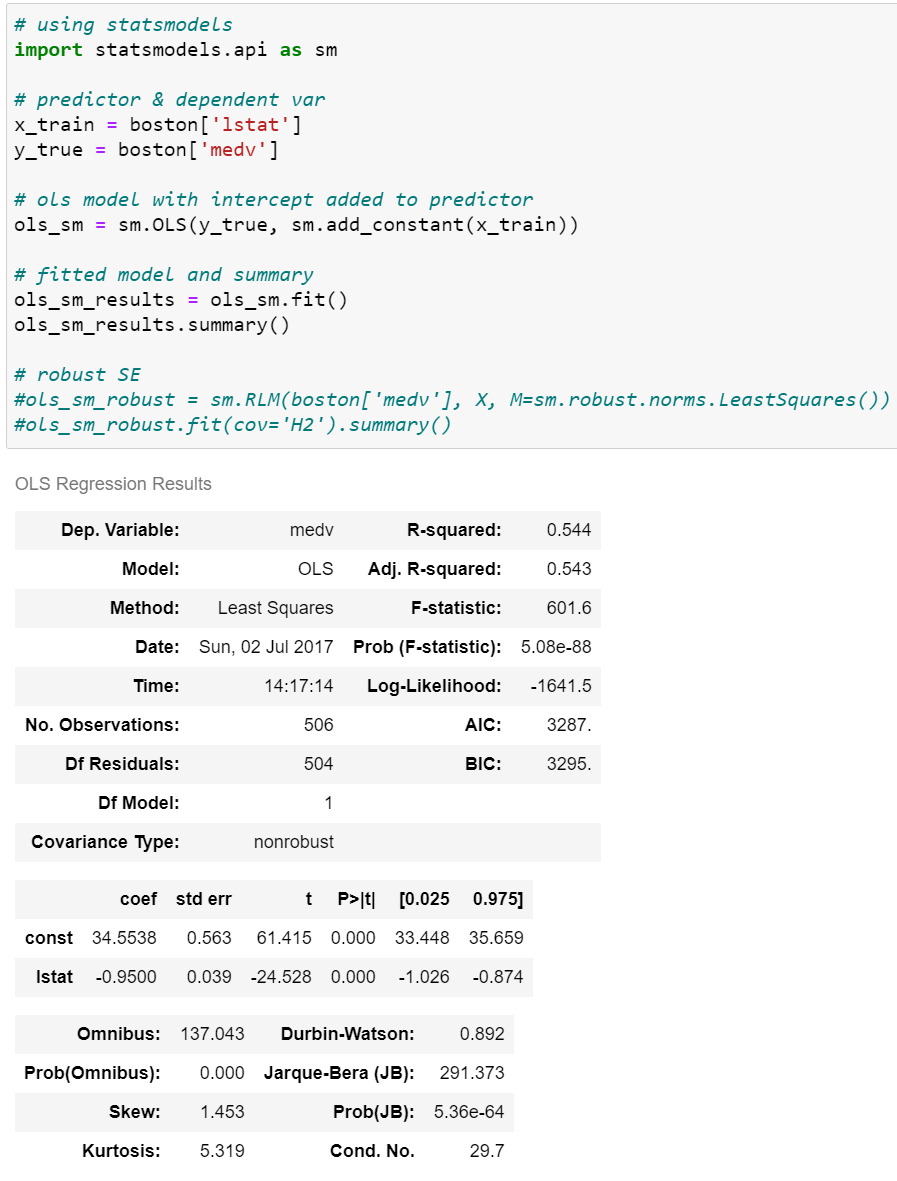


Predict using the linear model.

©copyright CIEL 2015 **42**

## 3.6.2 Simple Linear Regression

#### Using statsmodels



Ordinary least squares

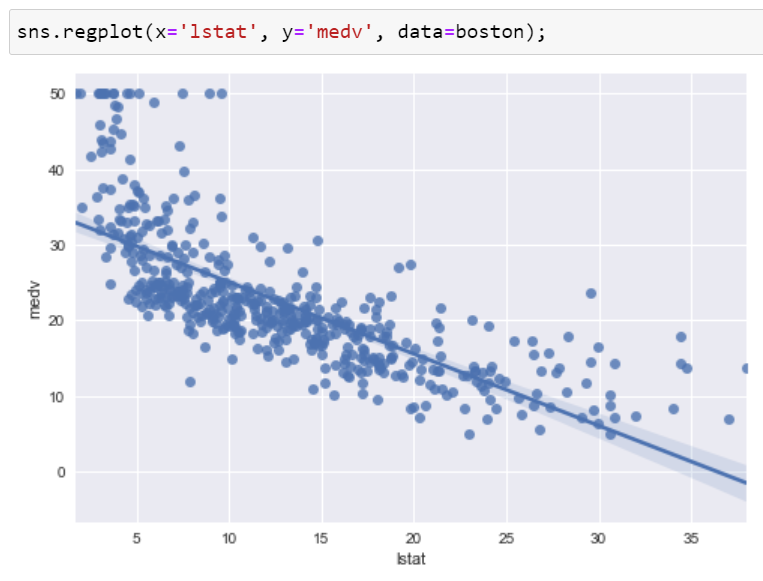
𝛽𝛽0 + 𝛽𝛽1𝑥𝑥

→ 𝛽𝛽0 × 1 + 𝛽𝛽1𝑥𝑥



Predicting responses

Plot data and a linear regression model fit.

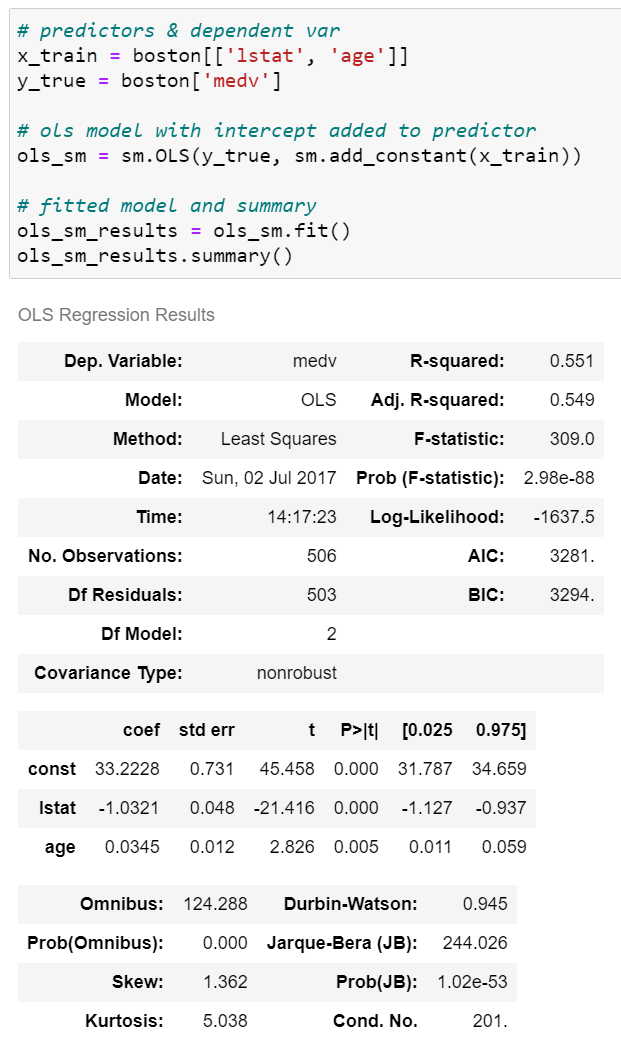


©copyright CIEL 2015 **43**

## 3.6.3 Multiple Linear Regression

#### Using statsmodels.api  Using statsmodels.formula.api

We have to add intercept

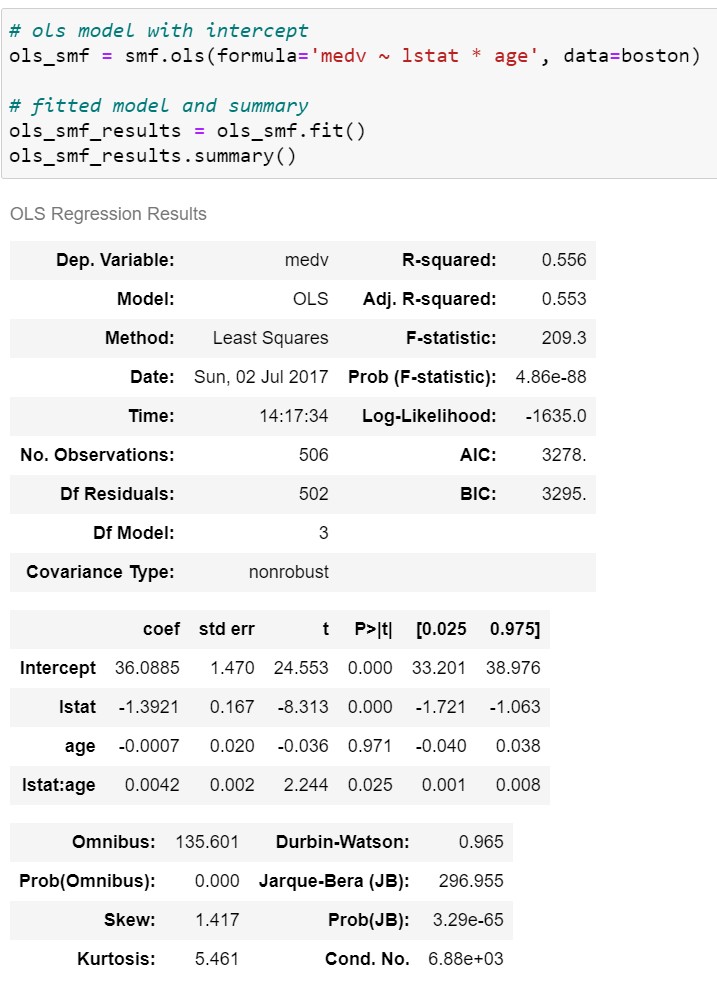


Automatically added intercept

©copyright CIEL 2015 **44**

## 3.6.4 Interaction Terms

#### Using statsmodels.formula.api



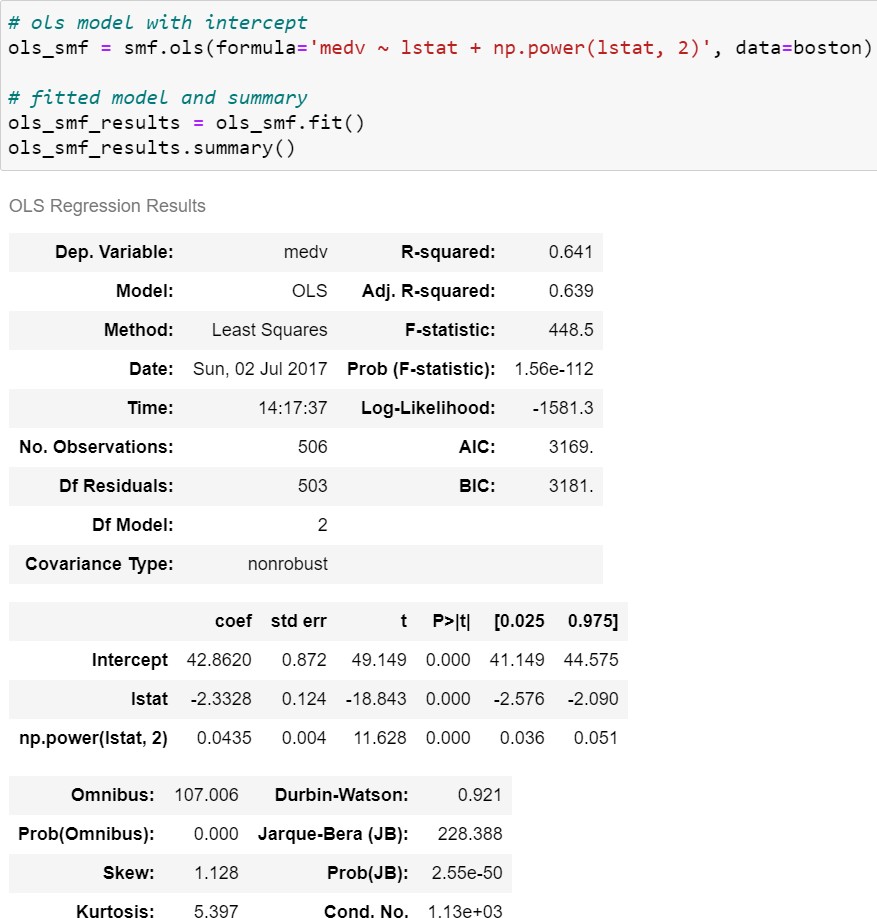
including the individual columns that were multiplied together

* + - Fitting models using R-style formulas

©copyright CIEL 2015 **45**

## 3.6.5 Non-linear Transformations of Predictors

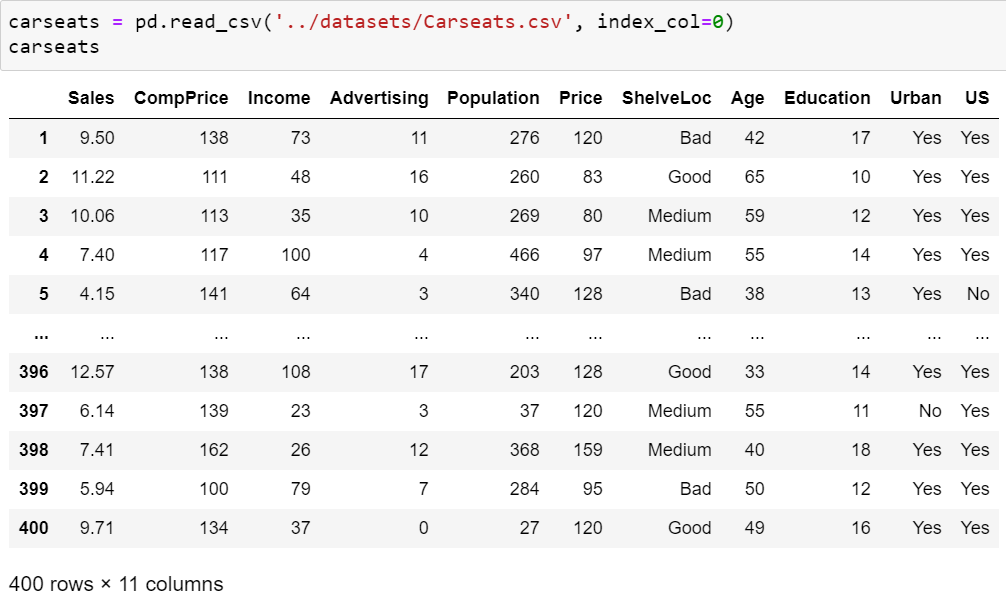
#### Using statsmodels.formula.api



©copyright CIEL 2015 **46**

## 3.6.6 Qualitative Predictors

#### Load dataset



Shelving location: space within a store in which the car seat is displayed

©copyright CIEL 2015 **47**

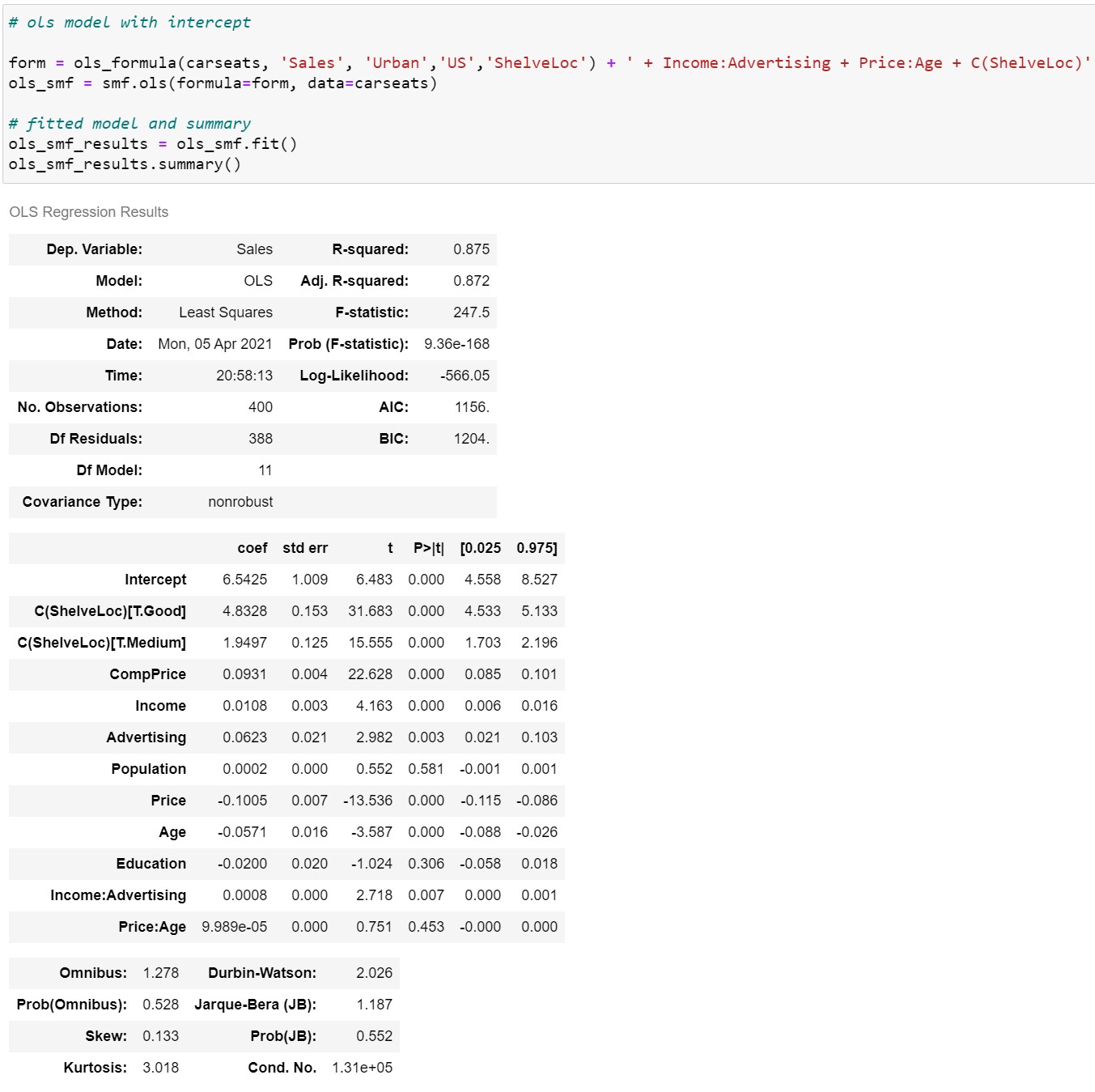
## 3.6.6 Qualitative Predictors

#### A user defined function for R style formula



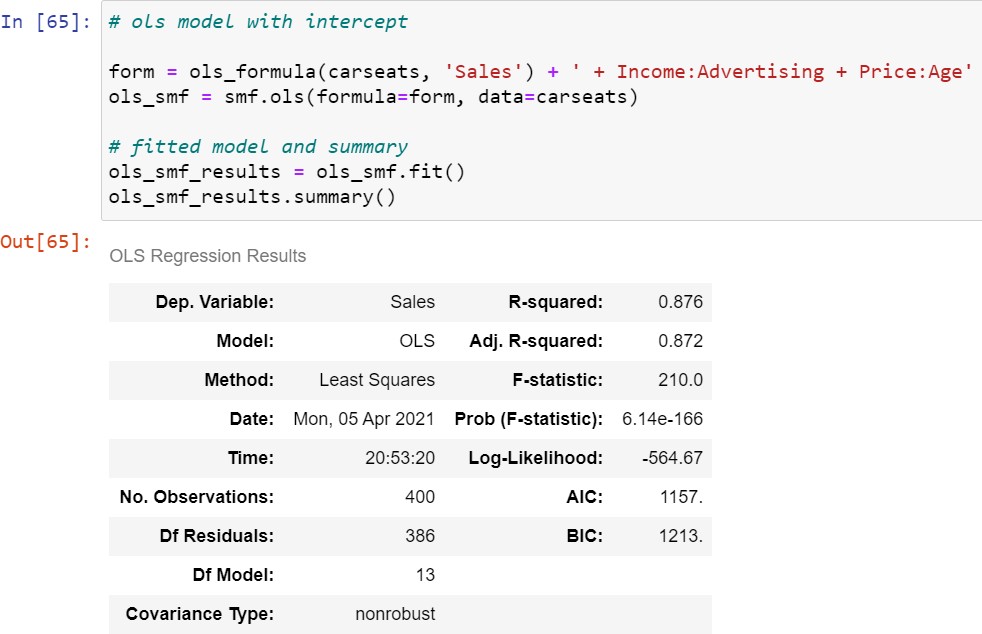
©copyright CIEL 2015 **48**

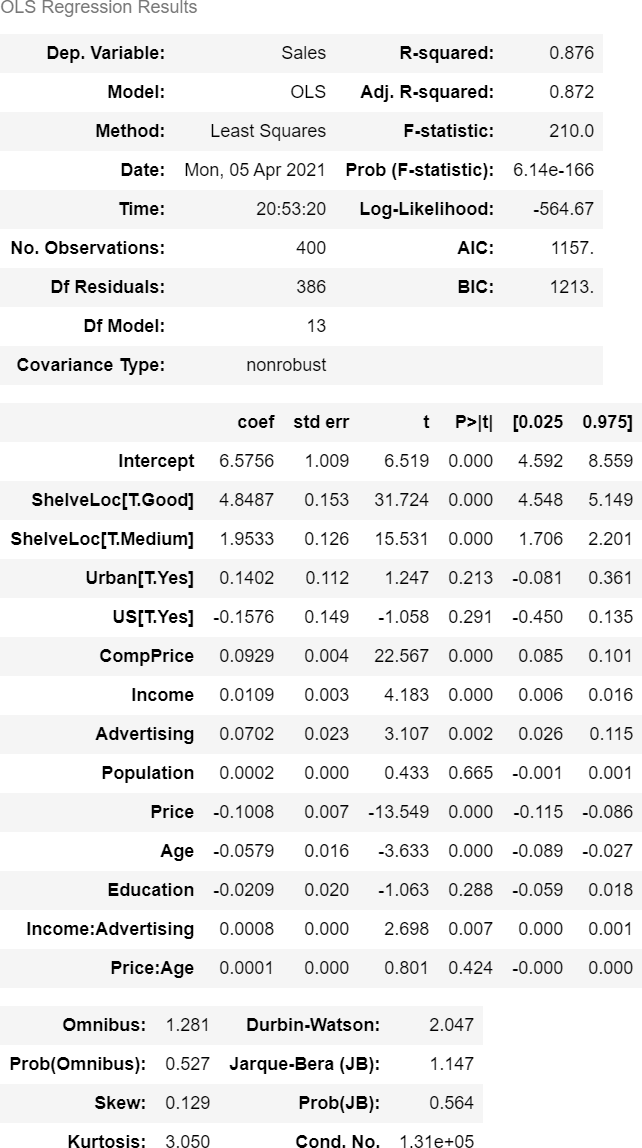
## [P] 3.6.6 Qualitative Predictors



©copyright CIEL 2015 **49**

## 3.6.6 Qualitative Predictors





©copyright CIEL 2015 **50**



**Summary & Next Class**

#### Multiple linear regression

* + Other considerations in regression model
  + Python lab
  + Summary & Next class

©copyright CIEL 2015 https://ciel.seoultech.ac.kr/

## Summary

#### Multiple linear regression

* + Relationship between response and predictors: using 𝐹𝐹-statistics
  + Important variables: forward & backward selection
  + Model fit: 𝑅𝑅2, RSE, and synergy or interaction effect
  + Predictions: confidence interval < prediction interval

#### Other considerations in regression model

* + Qualitative predictors: using dummy variables
  + Extension of linear model: interactions
  + Polynomial regression
  + Potential problems: outlier, high leverage, collinearity

#### Python lab

* + scikit-learn: sklearn.linear\_model.LinearRegression.fit
  + statsmodels.api.OLS.fit, statsmodels.formula.api.ols.fit

©copyright CIEL 2015 **52**

## Assignments

* eClass > Assignments
  + Upload 2 or 3 files (do not compress them)
* Python practices in today’s lecture
  + Upload a single ipynb file
  + Referring to the lecture slides marked with [P]
  + File name: “StudentID” + “\_AssignmentNo w/ 2 digits” + “\_1.ipynb”, e.g., **20211234\_02\_1.ipynb**
* Textbook exercise problems for today’s lecture
  + Conceptual
    - Solving the given problems, then upload your own solution (only docx/hwp formats, not pdf/jpg/bmp etc.)
    - Only include your answers (not need to describe problems)
    - File name: “StudentID” + “\_AssignmentNo w/ 2 digits” + “\_2.ipynb”, e.g., **20211234\_02\_2.docx**
  + Applied
    - Implement your Python code for the given problems, then upload another single ipynb file
    - File name: “StudentID” + “\_AssignmentNo w/ 2 digits” + “\_1.ipynb”, e.g., **20211234\_02\_3.ipynb**
* If not complying with the above policies, some penalty on assignment scores may be given.

©copyright CIEL 2015 **53**

## Course Schedule (Tentative)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Week** | **Topics** | **Note** | **Date (W)** | **Date (M)** |
| 1 | Orientation, Statistical Learning (Ch2) | Online | 03/03 | 03/08 |
| 2 | Statistical Learning (Ch2), Python Programming | Online | 03/10 | 03/15 |
| 3 | Probability & Statistics | Online | 03/17 | 03/22 |
| 4 | Probability & Statistics | Online | 03/24 | 03/29 |
| 5 | Linear Regression (Ch3) | Online | 03/31 | 04/05 |
| 6 | Linear Regression (Ch3) | Online | 04/07 | 04/12 |
| **7** | Classification (Ch4) | Online | 04/14 | 04/19 |
| 8 | **Midterm exam** | **7pm or Class hours (W1-W7)** | **04/21or26** | **04/21or26** |
| 9 | Resampling Methods (Ch5) | Online | 04/28 | 05/03 |
| 10 | Linear Model Selection and Regularization (Ch6) | Online | 05/05 | 05/10 |
| 11 | Moving Beyond Linearity (Ch7) | Online | 05/12 | 05/17 |
| 12 | Tree-Based Methods (Ch8) | Online | 05/19 | 05/24 |
| 13 | Support Vector Machines (Ch9) | Online | 05/26 | 05/31 |
| 14 | Unsupervised Learning (Ch10) | Online | 06/02 | 06/07 |
| 15 | **Final exam** | **7pm or Class hours (W9-W14)** | **06/09or14** | **06/09or14** |

©copyright CIEL 2015 **54**