Regression Week 1: Simple Linear Regression Assignment

Predicting House Prices (One feature)

In this notebook we will use data on house sales in King County, where Seattle is located, to predict house prices using simple (one feature) linear regression. You will:

- Use SArray and SFrame functions to compute important summary statistics
- Write a function to compute the Simple Linear Regression weights using the closed form solution
- Write a function to make predictions of the output given the input feature
- Turn the regression around to predict the input/feature given the output
- Compare two different models for predicting house prices

If you are doing the assignment with IPython Notebook

An IPython Notebook has been provided below to you for this assignment. This notebook contains the instructions, quiz questions and partially-completed code for you to use as well as some cells to test your code.

What you need to download

If you are using GraphLab Create:

- Download the King County House Sales data In SFrame format: kc_house_data.gl.zip (https://eventing.coursera.org/api/redirectStrict/VlCN2w4R28gwBfwYvch8wp1NDl1Knuo2AXf8BXWTH7a6_7V5-DxbY6-RqGWgX65c93salJg53BFqF0kPVy9e4A.5pCK9NQLw-gzeYiMI18pDQ.EN9RGaX5NYwseiuf_AMIMLdngaVTKm_WSjJ4VDCdj3-1c7ewt2qKQWCqAAWm51bwLadZd4UMX50ZyoNo4ZPMrHRnGwC8Yd-bI87y3I4Shr5Yyud5kMTLD7w66Ph4NGGvTIhBy-yPeaazrV7gh4GOUEmmuN5LWmNoNYv-sXJ5Of4U106-y9OD835nq6VtDrkYp7V0b1MHGuCZSa1KnaLjZIIpEhZWqT2yUM6avxOf65tCT8u8JyFiGFZGj2uAVYiwUDMnZ_Qzw0dmlRHGw095xoY1OcR-qMGgytGX1jTSKXJtJblRuNmAJtmNK9Lkh9IZEknVGyvG-p2yAPw0aCOgWVIacVYsrveYDzB85uUIP1puv0ADtgDWQE7baSxBDffPBS3BoSNeL4bC9yGLULEYxyP_aGs3zcEEpEJt5r-u2vPfzE8MuQ1dN3gJsrhqwXlv)
- Download the companion IPython Notebook: week-1-simple-regression-quiz-blank.ipynb (https://eventing.coursera.org/api/redirectStrict/jN_jGbCxNgQd8hxau-TrQAJwg-D3_2_44-

K8ihfB-4wOvmQQ9J1DexXr--AwRA-DyAB6AMSMXVkeyLW5fY0HNQ.RKEPufxsQKrsNHqKbVnwkQ.JJo0ZuiCCkHxIEqQlIIELXWhJmjkF9_-uVD7HqTBPZuMg-BFCGqlP13XqycItmWpQnNejuasmJMQABHQsRzGMqtXAZdVlyHDMCdH6meyZVWZXs90BcV_PBsTJdu7MlikhL1EqFFKfb4S7XGucoSRLEXXt9V9NQ2AOQi2rL4_wF-K0Kd9rWWhrfkoHg4AwxNa4nYl7O60nhik99i4vGbj_qiKP-78oXirGZ88z_u8W3nOgLbFpFiOf4duC45Ima9Qc288AFsCqeoPB5Y29CaMihBbB6keKyfp2m2g-vne4fHzpLJNFmRDDCfiHZnp5w9YBE188PnWpl4ge2o2D17WN90zgn20Ht1Q4YLewsGB11kzr6N8ap1PeObRPWfNwAtCRPkiQ4tdolK22JwySgfm-Ks9DD7CNMobAihbAxafqF4Rnev7zC3lN1FTUMMRcw6iUBjBaHYZM8MyVDn5BQVxY_qv5PhzDQ4QRfw038D_vYAgeNTnP8_5uoxr3T_JN0KN)

- Save both of these files in the same directory (where you are calling IPython notebook from) and unzip the data file.

If you are not using GraphLab Create:

- Download the King County House Sales data csv file: kc_house_data.csv (https://eventing.coursera.org/api/redirectStrict/fmES4TocNkHY5-EQiQTvGOf_oJK41xol9zCRw4ifEhfkDzjlOscs71u60dR5uFTM5psFgYFWxMNJ7ZcfaqSYRw.NO7njDik8Zh0nhtFBaiugg.tIJpPDjqWCu3OislXQGj-n7vH0J_Ut37JY5JtxRs2kJCmwVgNYFHEonnx6xi8Ieq3dF5Rogu65IA0KsoPSTcgBYSj4DqWwWSrhAxpvPbVYbHh_bBvHqfdpYezlJBvGNWyatD8AxRJGJHdu6fuI2PWElwvZlLglJI9bID07_TAbFu8dic6Hr3YjQfP2AK0FepJAZVDJJxNiWQGRsyZ-ohWwNfWs1Zdq3g7MaIcUQmTpGm063qhUSAFOcfG3vIPgmSOFTmZx3rqU809E8QsV2MNddYVFWMEQgFKCjPLdt1yxk0tjC-MMkUw-6AH6sBsZBD-qb5RR9ZWj-KT2NfF4LPrOZMUIOXAzauSiqCfMoG_f_0RFUGkRpLOpQP18F_pKLeFaoLITBoYgHQNBvME-a3fU3EQojMxjPvRpzQbZnjBhkXdPRFM0o01Q2vcVtSex0N)
- Download the King County House Sales training data csv file: kc_house_train_data.csv (https://eventing.coursera.org/api/redirectStrict/uHptNLlYMkdviF6ylpjLvl6qdr7KJ1G5i7LwRpQnm5XJ5_4QoXpA13YV4OtBKACGJ51M-516_Y6OX-iVyGyxlg.5rZOmhR57PklVcjYz4YNuw.7rOKiVP1Uu_SMv911KsQw-7CnXw1u9wE6MXJbo0jC8TXi3awD4Z1_W0N7ZI46Rr4H42D5DFgowefUdQ3GWZxx8IUaWJ3hwfuQqhneF_PDki7vqR9CoAWCm568DpQzsp5yNkYKj-VpIKccOmT09ukO23Zh1tSMjXa33HiQLTLxz2o4INEbZ8MbTKZax3lPJ651iOc831v08s0UCtr_TUiqY-E5MQdn-9E-_usvGszfuZvGt5KDiDH85ai2RnY42ndlKd7Tg818mFXCrdyRXl3HCtsDfpshVYxCRHN_ilPq2YVe07UOlwIrdDztQECW4v-bXVfr3NISDzzfLd4BdPt22FWCNNVURDH_52fIRmOzbWRVbnjzkNreIFscpql_KAo3lvbvF2fZE2lE6YRhXnfsszEIVo8YhrMP2W6vU7Cuo-PHYsSZiw3a7fxmOq41p6U5H449Hbi5zntXsbGfD1-XA)
- Download the King County House Sales testing data csv file: kc_house_test_data.csv (https://eventing.coursera.org/api/redirectStrict/uY8CJuE3hfumEYKgnCli4L4TIcMvqn_OjkkRTVW1DBgVt16AZ2mphpczM8ekxVlCBbjl6kqLX5TyEJ7v19OY4w.FuAV359GSH2Zn1-

MOyAlYQ.n1jE3Foq6nFhv1CSfUvvtJrWZLR2dFeF4u_xfAHJz_c4DEgd9tLQ4K1XUOlCu0pKJWUCm1IeYv_otdT6hBXcQxKw6QRg6GkavotgmAUkRQPpWfNBsooPsUt9O9I-l0FHsKCo5B5d-EQwIvQFs_sVAbhLnxHppN7L0QT12ApiO4J-OrFLbUYOeXhOGQut0e_6lV3FYd46OiwHVnEgOPVOKvjOwjAx69usgJJjRQm4r_QuY9PZze7kRpWYmowJ9YzD-TDM8WBTtohphrFAWf6k_3mF_JGfxvlY1pnHTYSUGpgKeS2l9-Fn2i-ZspRr9iaFNPxsGSTwkel2_RzxzCpmYhmSoGSfovcr3PwMPHqvMYnIk7CUb6szcFluZLKATnQclsPg1bCESojaBzq-aB2IoFGKE-f7GzhJ29pP-e4GooclMjWomIBqdBsCA04qA8y5Wp917P3Ioe2ZVryfz83QiA)

- **IMPORTANT: use the following types for columns when importing the csv files. Otherwise, they may not be imported correctly: [str, str, float, float, float, float, int, str, int, int, int, int, int, int, int, str, float, float, float, float]. If your tool of choice requires a dictionary of types for importing csv files (e.g. Pandas), use:**

```
dtype_dict = {'bathrooms':float, 'waterfront':int, 'sqft_above':int, 'sqft_living15':float, 'grade':int, 'yr_renovated':int, 'price':float, 'bedrooms':float, 'zipcode':str, 'long':float, 'sqft_lot15':float, 'sqft_living':float, 'floors':str, 'condition':int, 'lat':float, 'date':str, 'sqft_basement':int, 'yr_built':int, 'id':str, 'sqft_lot':int, 'view':int}
```

Useful resources

You may need to install the software tools or use the free Amazon EC2 machine. Instructions for both options are provided in the reading for Module 1.

If instead you are using other tools to do your homework

You are welcome, however, to write your own code and use any other libraries, like Pandas or R, to help you in the process. If you would like to take this path, follow the instructions below.

1. If you are using SFrame, import graphlab and load in the house data, otherwise you can also download the csv. (Note that we will be using the training and testing csv files provided). e.g in python with SFrames:

```
sales = graphlab.SFrame('kc_house_data.gl/')
```

2. Split data into 80% training and 20% test data. Using SFrame, use this command to set the same seed for everyone. e.g. in python with SFrames:

```
train_data,test_data = sales.random_split(.8,seed=0)
```

For those students not using graphlab please download the training and testing data csv files.

From now on we will train the models using train_data. It will be important that we use the same split here to ensure the results are the same.

3. Write a generic function that accepts a column of data (e.g, an SArray) 'input_feature' and another column 'output' and returns the Simple Linear Regression parameters 'intercept' and 'slope'. Use the closed form solution from lecture to calculate the slope and intercept. e.g. in python:

```
def simple_linear_regression(input_feature, output):
    [your code here]
return(intercept, slope)
```

4. Use your function to calculate the estimated slope and intercept on the training data to predict 'price' given 'sqft_living'. e.g. in python with SFrames using:

```
input_feature = train_data['sqft_living']
output = train_data['price']
```

save the value of the slope and intercept for later (you might want to call them e.g. squarfeet_slope, and squarefeet_intercept)

5. Write a function that accepts a column of data 'input_feature', the 'slope', and the 'intercept' you learned, and returns an a column of predictions 'predicted_output' for each entry in the input column. e.g. in python:

```
def get_regression_predictions(input_feature, intercept, slope)
    [your code here]
return(predicted_output)
```

6. Quiz Question: Using your Slope and Intercept from (4), What is the predicted price for a house with 2650 sqft?

7. Write a function that accepts column of data: 'input_feature', and 'output' and the regression parameters 'slope' and 'intercept' and outputs the Residual Sum of Squares (RSS). e.g. in python:

```
def get_residual_sum_of_squares(input_feature, output, intercept,slope):
    [your code here]
return(RSS)
```

Recall that the RSS is the sum of the squares of the prediction errors (difference between output and prediction).

8. Quiz Question: According to this function and the slope and intercept from (4) What is the RSS for the simple linear regression using squarefeet to predict prices on TRAINING data?

9. Note that although we estimated the regression slope and intercept in order to predict the output from the input, since this is a simple linear relationship with only two variables we can invert the linear function to estimate the input given the output!

Write a function that accept a column of data:'output' and the regression parameters 'slope' and 'intercept' and outputs the column of data: 'estimated_input'. Do this by solving the linear function output = intercept + slope*input for the 'input' variable (i.e. 'input' should be on one side of the equals sign by itself). e.g. in python:

```
def inverse_regression_predictions(output, intercept, slope):
    [your code here]
return(estimated_input)
```

10. Quiz Question: According to this function and the regression slope and intercept from (3) what is the estimated square-feet for a house costing $800,000?

11. Instead of using 'sqft_living' to estimate prices we could use 'bedrooms' (a count of the number of bedrooms in the house) to estimate prices. Using your function from (3) calculate the Simple Linear Regression slope and intercept for estimating price based on bedrooms. Save this slope and intercept for later (you might want to call them e.g. bedroom_slope, bedroom_intercept).

12. Now that we have 2 different models compute the RSS from BOTH models on TEST data.

13. Quiz Question: Which model (square feet or bedrooms) has lowest RSS on TEST data? Think about why this might be the case.