Week03

Assignment

Assignment 02

- Build linear regression model
  - Among numeric variables, select input variables
    - Describe reasons for variable selection
  - You can apply variable transformation
    - Describe reasons of variable transformation
  - You can discard some rows satisfying specific conditions
    - Specify conditions
- Summarize the process and result using Power Point
  - Some students have to create video clip to explain their results

- Submit both slide and python code

Linear Regression

Trained Model

OLS Regression Results

Model

Dep. Variable:	price	R-squared:	0.598
Model:	OLS	Adj. R-squared:	0.598
Method:	Least Squares	F-statistic:	2923.
Date:	Tue, 08 Sep 2020	Prob (F-statistic):	0.00
Time:	08:46:44	Log-Likelihood:	-2.9775e+05
No. Observations:	21613	AIC:	5.955e+05
Df Residuals:	21601	BIC:	5.956e+05
Df Model:	11		
Covariance Type:	nonrobust		

Coefficients

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.418e+06	1.38e+05	46.627	0.000	6.15e+06	6.69e+06
bedrooms	-5.813e+04	2150.015	-27.038	0.000	-6.23e+04	-5.39e+04
bathrooms	6.615e+04	3737.401	17.701	0.000	5.88e+04	7.35e+04
sqft_lot	0.0371	0.055	0.671	0.502	-0.071	0.145
floors	5.498e+04	4020.903	13.673	0.000	4.71e+04	6.29e+04
waterfront	7.247e+05	1.86e+04	39.027	0.000	6.88e+05	7.61e+05
sqft_above	239.6824	3.895	61.538	0.000	232.048	247.317
sqft_basement	243.7353	4.812	50.654	0.000	234.304	253.167
yr_built	-3338.9292	71.492	-46.703	0.000	-3479.059	-3198.799
yr_renovated	11.9013	4.156	2.864	0.004	3.756	20.047
sqft_living15	90.4224	3.679	24.581	0.000	83.212	97.633
sqft_lot15	-0.7360	0.084	-8.731	0.000	-0.901	-0.571

Residuals

Omnibus:	14160.528	Durbin-Watson:	1.981
Prob(Omnibus):	0.000	Jarque-Bera (JB):	606691.177
Skew:	2.579	Prob(JB):	0.00
Kurtosis:	28.438	Cond. No.	4.40e+06

Trained Model

Model

OLS Regression Results

Dep. Variable:	price	R-squared:	0.598
Model:	OLS	Adj. R-squared:	0.598
Method:	Least Squares	F-statistic:	2923.
Date:	Tue, 08 Sep 2020	Prob (F-statistic):	0.00
Time:	08:46:44	Log-Likelihood:	-2.9775e+05
No. Observations:	21613	AIC:	5.955e+05
Df Residuals:	21601	BIC:	5.956e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.418e+06	1.38e+05	46.627	0.000	6.15e+06	6.69e+06
bedrooms	-5.813e+04	2150.015	-27.038	0.000	-6.23e+04	-5.39e+04
bathrooms	6.615e+04	3737.401	17.701	0.000	5.88e+04	7.35e+04
sqft_lot	0.0371	0.055	0.671	0.502	-0.071	0.145
floors	5.498e+04	4020.903	13.673	0.000	4.71e+04	6.29e+04
waterfront	7.247e+05	1.86e+04	39.027	0.000	6.88e+05	7.61e+05
sqft_above	239.6824	3.895	61.538	0.000	232.048	247.317
sqft_basement	243.7353	4.812	50.654	0.000	234.304	253.167
yr_built	-3338.9292	71.492	-46.703	0.000	-3479.059	-3198.799
yr_renovated	11.9013	4.156	2.864	0.004	3.756	20.047
sqft_living15	90.4224	3.679	24.581	0.000	83.212	97.633
sqft_lot15	-0.7360	0.084	-8.731	0.000	-0.901	-0.571

Omnibus:	14160.528	Durbin-Watson:	1.981
Prob(Omnibus):	0.000	Jarque-Bera (JB):	606691.177
Skew:	2.579	Prob(JB):	0.00
Kurtosis:	28.438	Cond. No.	4.40e+06

F-test

- F-test for general regression models
  - Check overall significance of regression models
    - Whether the regression model is significant for predicting a target
  - Hypothesis

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

$$H_1: \text{not all } \beta_i (i = 1,2,\cdots,p) \text{ equal zero}$$

  - Test statistic

$$F^* = MSR/MSE$$

    - F follows F-distribution with $(p, n-p-1)$ degree of freedom
  - Decision rule

$$\text{If } F^* \leq F(1-\alpha; p, n-p-1), \text{ conclude } H_0$$

$$\text{If } F^* > F(1-\alpha; p, n-p-1), \text{ conclude } H_1$$

    - α: significance level

Sum of Square

- Total variance: the total sum of squares

$$SST = \sum_i (y_i - \bar{y})^2$$

- Explained variance: the regression sum of squares, also called the explained sum of squares

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

- Residual variance: the sum of squares errors, also called the residual sum of squares

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

- Relationship among three values

$$SST = SSR + SSE$$

Test of Model Significance

- ANOVA table for multiple regression model with p input variables

Factor	Sum of square	Degree of freedom	Mean square	F-value	p-value
Model	SSR	p	$MSR = SSR/p$	$F_0 = \mathrm{MSR}/\mathrm{MSE}$	$P\{F_{p,n-p-1} > F_0\}$
Residual	SSE	$n-p-1$	$MSE = SSE/(n-p-1)$		
Total	SST	$n-1$			

  - Analysis of Variance (ANOVA)

R^2

- Statistical measures for goodness-of-fit
  - $R^2 \;\; (0 \le R^2 \le 1)$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Adjusted R^2

$$R^2{}_{adj} = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}} = 1 - \left(\frac{n-1}{n-p-1}\right)(1 - R^2)$$

Depend on the number of input variables

  - Penalty on the number of input variable by $n - p - 1$
  - Adjusted R^2 may actually become smaller when another input variable is introduced into the model

Likelihood of Linear Regression Model

- Consider linear regression model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

- Joint density of the independent random responses $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ evaluated at observations(true), $\mathbf{y} = (y_1, \dots, y_n)^T$

$$f(\mathbf{y}; \boldsymbol{\beta})$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_1 - \beta_0 - \beta_1 x_{11} - \cdots - \beta_p x_{p1})^2}{2\sigma^2}} \times \cdots \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_n - \beta_0 - \beta_1 x_{1n} - \cdots - \beta_p x_{pn})^2}{2\sigma^2}}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi})^2}$$

$\Downarrow$ **Maximum likelihood estimation**

$$\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$$

AIC, BIC

- Akaike information criterion (AIC) and Bayesian Information Criterion (BIC)
  - Estimators of in-sample prediction error and thereby relative quality of statistical models for a given set of data
  - The model with the lowest AIC or BIC is preferred

$$AIC = -2 \log \mathcal{L} + 2p$$
$$BIC = -2 \log \mathcal{L} + \log(n) \cdot p$$

    - $\mathcal{L}$: likelihood of the model
    - p: the number of estimated parameters in the model

Trained Model - Coefficients

Coefficients

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.598
Model:                            OLS   Adj. R-squared:                  0.598
Method:                 Least Squares   F-statistic:                     2923.
Date:                Tue, 08 Sep 2020   Prob (F-statistic):               0.00
Time:                        08:46:44   Log-Likelihood:            -2.9775e+05
No. Observations:               21613   AIC:                         5.955e+05
Df Residuals:                   21601   BIC:                         5.956e+05
Df Model:                          11
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          6.418e+06   1.38e+05     46.627      0.000    6.15e+06    6.69e+06
bedrooms      -5.813e+04   2150.015    -27.038      0.000   -6.23e+04   -5.39e+04
bathrooms      6.615e+04   3737.401     17.701      0.000    5.88e+04    7.35e+04
sqft_lot          0.0371      0.055      0.671      0.502      -0.071       0.145
floors         5.498e+04   4020.903     13.673      0.000    4.71e+04    6.29e+04
waterfront     7.247e+05   1.86e+04     39.027      0.000    6.88e+05    7.61e+05
sqft_above      239.6824      3.895     61.538      0.000     232.048     247.317
sqft_basement   243.7353      4.812     50.654      0.000     234.304     253.167
yr_built      -3338.9292     71.492    -46.703      0.000   -3479.059   -3198.799
yr_renovated     11.9013      4.156      2.864      0.004       3.756      20.047
sqft_living15    90.4224      3.679     24.581      0.000      83.212      97.633
sqft_lot15       -0.7360      0.084     -8.731      0.000      -0.901      -0.571
==============================================================================
Omnibus:                    14160.528   Durbin-Watson:                   1.981
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           606691.177
Skew:                           2.579   Prob(JB):                         0.00
Kurtosis:                      28.438   Cond. No.                     4.40e+06
==============================================================================
```

Test Concerning Regression Coefficients

- Test for $\beta_j (j = 0,1,2,\cdots,p)$
  - Hypothesis

$$H_0: \beta_j = 0$$
$$H_1: \beta_j \neq 0$$

  - Test statistic

$$t_j = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)}$$

    - $se^2(\widehat{\boldsymbol{\beta}}) = MSE(\mathbf{X}^T\mathbf{X})^{-1} \rightarrow se^2(\widehat{\beta}_j) = [MSE(\mathbf{X}^T\mathbf{X})^{-1}]_{j,j}$

  - Decision rule

If $|t_j| \leq t\left(1 - \frac{\alpha}{2}; n - p - 1\right)$, conclude H_0

If $|t_j| > t\left(1 - \frac{\alpha}{2}; n - p - 1\right)$, conclude H_1

Trained Model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.598
Model:                            OLS   Adj. R-squared:                  0.598
Method:                 Least Squares   F-statistic:                     2923.
Date:                Tue, 08 Sep 2020   Prob (F-statistic):               0.00
Time:                        08:46:44   Log-Likelihood:            -2.9775e+05
No. Observations:               21613   AIC:                         5.955e+05
Df Residuals:                   21601   BIC:                         5.956e+05
Df Model:                          11
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          6.418e+06   1.38e+05     46.627      0.000    6.15e+06    6.69e+06
bedrooms      -5.813e+04   2150.015    -27.038      0.000   -6.23e+04   -5.39e+04
bathrooms      6.615e+04   3737.401     17.701      0.000    5.88e+04    7.35e+04
sqft_lot          0.0371      0.055      0.671      0.502      -0.071       0.145
floors         5.498e+04   4020.903     13.673      0.000    4.71e+04    6.29e+04
waterfront     7.247e+05   1.86e+04     39.027      0.000    6.88e+05    7.61e+05
sqft_above      239.6824      3.895     61.538      0.000     232.048     247.317
sqft_basement   243.7353      4.812     50.654      0.000     234.304     253.167
yr_built      -3338.9292     71.492    -46.703      0.000   -3479.059   -3198.799
yr_renovated     11.9013      4.156      2.864      0.004       3.756      20.047
sqft_living15    90.4224      3.679     24.581      0.000      83.212      97.633
sqft_lot15       -0.7360      0.084     -8.731      0.000      -0.901      -0.571
==============================================================================
Omnibus:                    14160.528   Durbin-Watson:                   1.981
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           606691.177
Skew:                           2.579   Prob(JB):                         0.00
Kurtosis:                      28.438   Cond. No.                     4.40e+06
==============================================================================
```

Residuals

Residual Analysis for Linear Regression

- Jarque-Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution
  - Test statistic

$$JB = \frac{n-k}{6}\left(S^2 + \frac{1}{4}(C-3)^2\right)$$

    - $S = \dfrac{\frac{1}{n}\Sigma_{i=1}^{n}(x_i-\bar{x})^3}{\left(\frac{1}{n}\Sigma_{i=1}^{n}(x_i-\bar{x})^2\right)^{\frac{3}{2}}}$: sample skewness
    - $C = \dfrac{\frac{1}{n}\Sigma_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\Sigma_{i=1}^{n}(x_i-\bar{x})^2\right)^{\frac{4}{2}}}$: sample kurtosis
    - k: the number of input variables
  - If the data comes from a normal distribution, JB statistic asymptotically has a chi-squared distribution with two degrees of freedom

$$H_0: S = C - 3 = 0$$

Residual Analysis for Linear Regression

- Durbin-Watson statistic is a test statistic used to detect the presence of autocorrelation at lag 1 in the residuals (prediction errors) from a regression analysis
  - It is used for time series data
  - If e_t is the residual given by $e_t = \rho e_{t-1} + \nu_t$
    - t is a timestamp
    - Hypothesis

$$H_0: \rho = 0$$
$$H_a: \rho \neq 0$$

    - Test statistic

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

  - Result
    - When $d = 2$, no autocorrelation
    - When $d < 2$, positive serial correlation
    - When $d > 2$, negative serial correlation

Residual Analysis for Linear Regression

- Omnibus test is a statistical test for normality
$$Z = Z(S)^2 + Z(C)^2 \sim \chi(2)$$
  - $Z(S)$ is z-score by test of skewness and $Z(C)$ is z-score by test of kurtosis
- Tests of Skewness (S)
$$Y = S\sqrt{\frac{(n+1)(n+3)}{6(n-2)}}$$
$$\beta_2(S) = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$$
$$W^2 = -1 + \sqrt{2\beta_2(S) - 2}$$
$$\delta = \frac{1}{\sqrt{\ln W}}$$
$$\alpha = \sqrt{\frac{2}{W^2 - 1}}$$
$$\rightarrow Z(S) = \delta \ln\left(\frac{Y}{\alpha} + \sqrt{\left(\frac{Y}{\alpha}\right)^2 + 1}\right)$$

Residual Analysis for Linear Regression

- Omnibus test is a statistical test for normality

$$Z = Z(S)^2 + Z(C)^2 \sim \chi(2)$$

  - $Z(S)$ is z-score by test of skewness and $Z(C)$ is z-score by test of kurtosis

  - Tests of kurtosis (C)

$$E(C) = \frac{3(n-1)}{n+1}$$

$$var(C) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

$$x = (C - E(C))/\sqrt{var(C)}$$

$$\sqrt{\beta_1(C)} = \frac{6(n^2-5n+2)}{(n+7)(n+9)}\sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}}$$

$$A = 6 + \frac{8}{\sqrt{\beta_1(C)}}\left[\frac{2}{\sqrt{\beta_1(C)}} + \sqrt{\left(1 + \frac{4}{\sqrt{\beta_1(C)}}\right)}\right]$$

$$\rightarrow Z(C) = \frac{\left(1-\frac{2}{9A}\right) - \left[\frac{1-\frac{2}{A}}{1+x\sqrt{\frac{2}{A-4}}}\right]^{\frac{1}{3}}}{\sqrt{2/(9A)}}$$

Handling Categorical Variables

Handle Categorical Variables

- Categorical variables of the dataset, House Sales Prices in King County
  - waterfront is binary variable
  - 'view', 'condition', 'grade' are ordinal variables (categorical variables)
  - 'zipcode' is a nominal variable

Regression with Categorical Variables

- Dummy variable
  - Also known as an indicator variable or binary variable
  - Take the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome

- Categorical variables with k categories → convert to vector represented by k binary variables using one-hot encoding
  - For binary variable

Original variable	→	$dummy_1$	$dummy_2$
Female		1	0
Male		0	1

- Actually $\sum_i x_i = 1$ should be satisfied → $k-1$ binary variables are necessary

Original variable	→	$dummy_1$	$dummy_2$
Female		1	0
Male		0	1

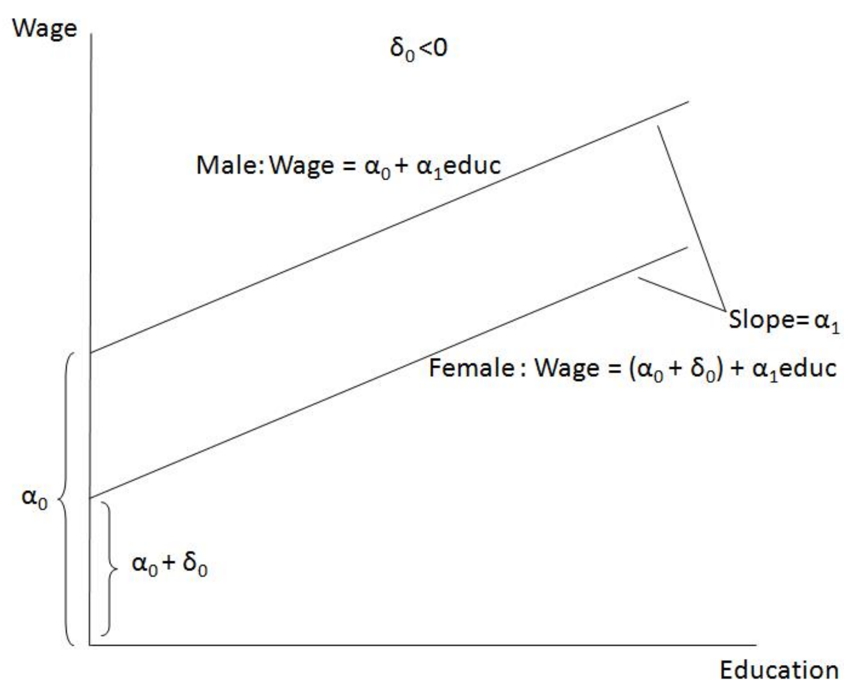
Regression with Categorical Variables

- Instead of categorical variables, dummy variables are used in regression function
  - The problem to predict wage by education and sex(female/male)

$$wage = \alpha_0 + \alpha_1 \cdot educ + \alpha_2 \cdot sex$$

    - Because sex is categorical variable create a dummy variable $dummy_f$ (female is 1 and male is 0)

$$wage = \alpha_0 + \alpha_1 \cdot educ + \delta_0 \cdot dummy_f$$



Regression with Categorical Variables

- Example
  - There are two categorical variables

	x_1	x_2
# of categories	3	4

	$dummy_{11}$	$dummy_{12}$	$dummy_{13}$
$x_1 = 1$	1	0	0
$x_1 = 2$	0	1	0
$x_1 = 3$	0	0	1

	$dummy_{21}$	$dummy_{22}$	$dummy_{23}$	$dummy_{24}$
$x_1 = 1$	1	0	0	0
$x_2 = 2$	0	1	0	0
$x_3 = 3$	0	0	1	0
$x_4 = 4$	0	0	0	1

Programming Exercise

Create Dummy Variable

- Pandas package provides function to create dummy variable
  - Use “get_dummies”

```
import pandas as pd

dummy1=pd.get_dummies(data['var'],prefix='var')
```

- It creates k binary variables if the categorical variable has k categories
- To reduce the number of variables in final train set, you can take the first $k-1$ dummy variables using drop_first option

Create Dummy Variable

- Load example data

```
import pandas as pd

salary=pd.read_csv('https://drive.google.com/uc?export=download&id=1kkAZzL8uRSak8gM-0iqMMAFQJTfnyGuh')
```

  - rank, discipline, and sex are categorical variables

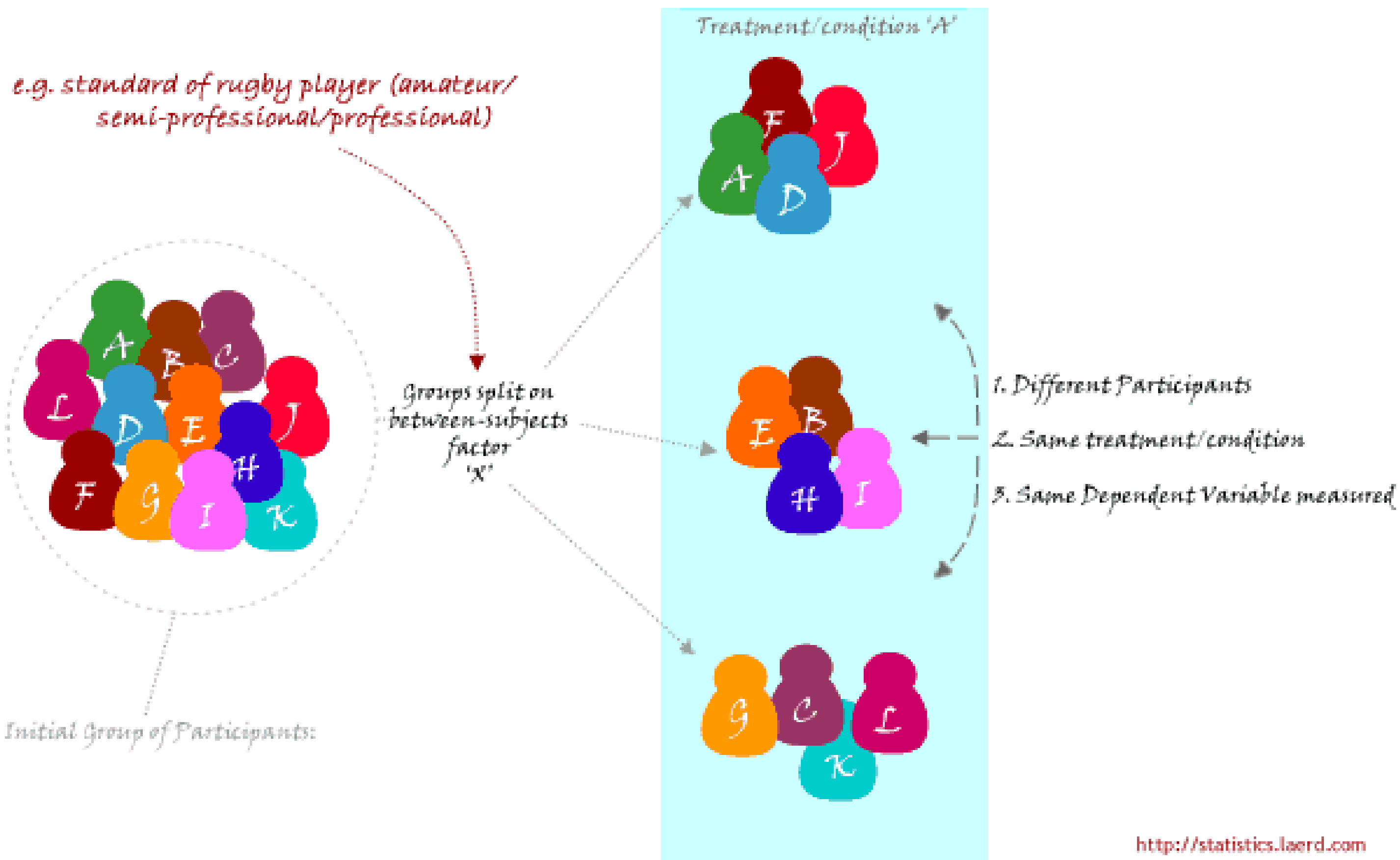
- Question
  - Calculate mean values of salary according to rank, discipline and sex

Build a Regression Model

- Question
  - Check relevancy or significance of categorical variables for predicting salary

One-way ANOVA

- One-way analysis of variance(ANOVA)
  - It is used to determine whether there are any statistically significant differences between the means of two or more independent (unrelated) groups



One-way ANOVA

- Hypothesis

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_a: The\ mean\ of\ at\ least\ group\ is\ different$$

- Assumptions
  - Normality - That each sample is taken from a normally distributed population
  - Sample independence - that each sample has been drawn independently of the other samples
  - Variance Equality - That the variance of data in the different groups should be the same

One-way ANOVA

Source	Sums of squares	Degrees of freedom	Mean square	F
Between Samples	$\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$	$k-1$	$\frac{SSB}{k-1}$	$\frac{MSB}{MSW}$
Within Samples	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$	$n-k$	$\frac{SSW}{n-k}$	
Total	$\sum_{i,j}(x_{ij} - \bar{x})^2$	$n-1$		

Build a Model: Linear Regression

- Use the Statsmodels package

```
import statsmodels.api as sm
model=sm.OLS(y, X)
result=model.fit()
result.summary()
```

  - Estimated parameters

```
result.params
```

  - Prediction

```
y_pred=result.predict(X)
```

  - Residual of training set

```
result.resid
```

Build a Model: Linear Regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 salary   R-squared:                       0.455
Model:                            OLS   Adj. R-squared:                  0.446
Method:                 Least Squares   F-statistic:                     54.20
Date:                Mon, 27 Jul 2020   Prob (F-statistic):           1.79e-48
Time:                        13:18:54   Log-Likelihood:                -4538.9
No. Observations:                 397   AIC:                             9092.
Df Residuals:                     390   BIC:                             9120.
Df Model:                           6
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          7.886e+04   4990.326     15.803      0.000    6.91e+04    8.87e+04
yrs.since.phd   535.0583    240.994      2.220      0.027      61.248    1008.869
yrs.service    -489.5157    211.938     -2.310      0.021    -906.199     -72.833
rank_AsstProf -1.291e+04   4145.278     -3.114      0.002   -2.11e+04   -4757.700
rank_Prof      3.216e+04   3540.647      9.083      0.000    2.52e+04    3.91e+04
discipline_B   1.442e+04   2342.875      6.154      0.000    9811.380     1.9e+04
sex_Male       4783.4928   3858.668      1.240      0.216   -2802.901    1.24e+04
==============================================================================
Omnibus:                       46.385   Durbin-Watson:                   1.919
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               82.047
Skew:                           0.699   Prob(JB):                     1.53e-18
Kurtosis:                       4.733   Cond. No.                         183.
==============================================================================
```

Build a Model: Linear Regression

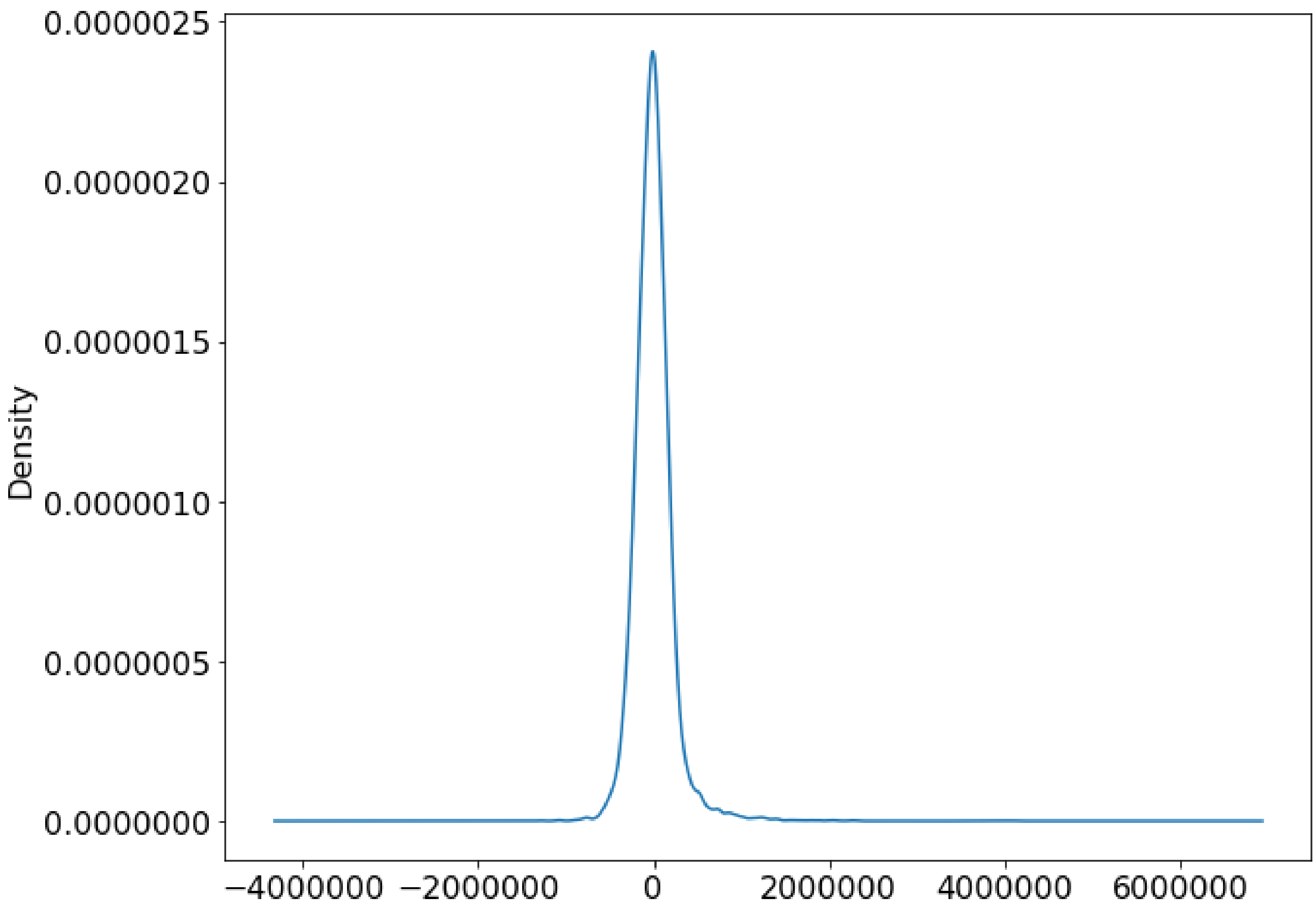
- House Sales Prices in King County
  - Explanatory variables
    - Use numeric variables including ‘waterfront’
    - Check multicollinearity using VIF

Build a Model: Linear Regression

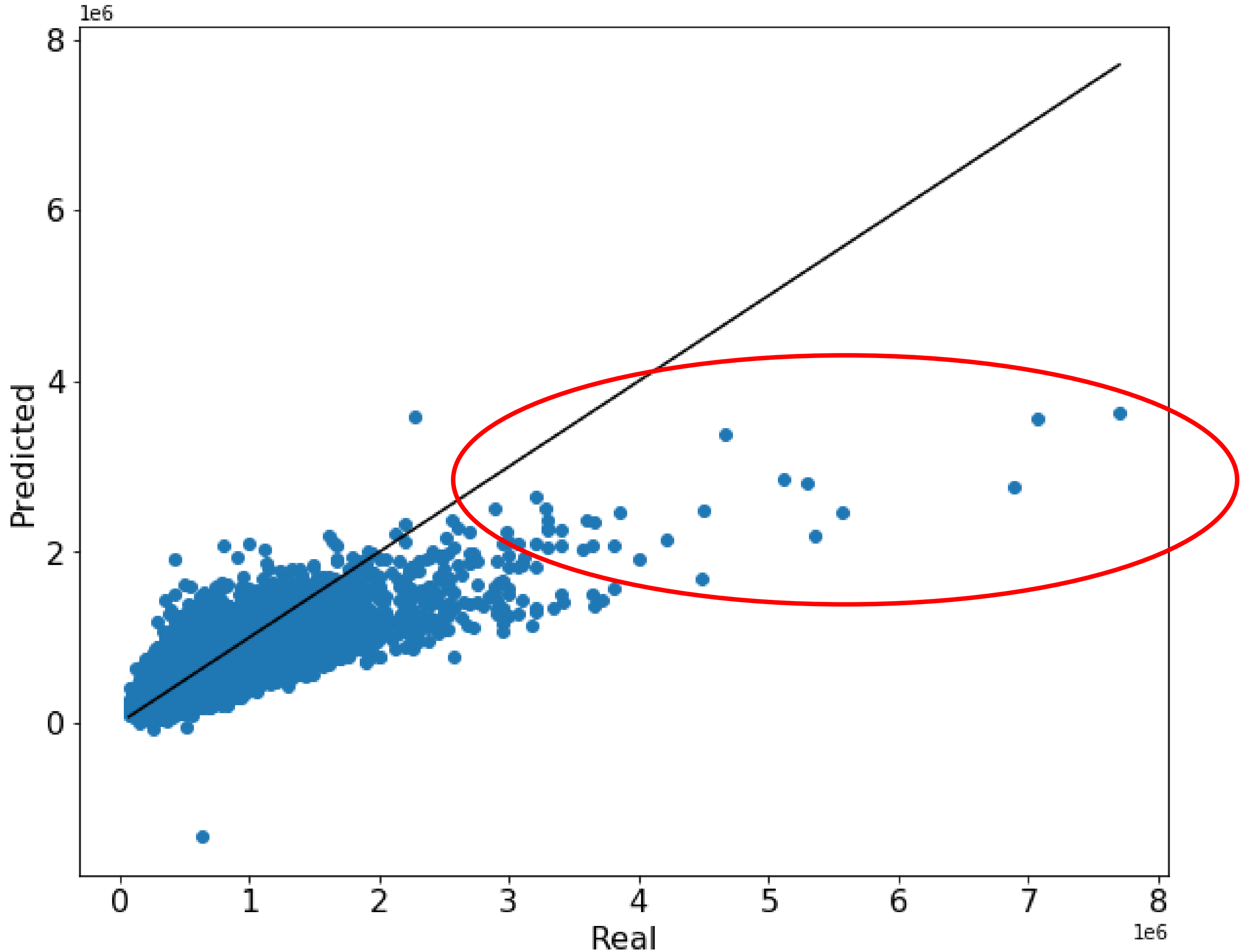
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.598
Model:                            OLS   Adj. R-squared:                  0.598
Method:                 Least Squares   F-statistic:                     2923.
Date:                Mon, 27 Jul 2020   Prob (F-statistic):               0.00
Time:                        12:16:59   Log-Likelihood:            -2.9775e+05
No. Observations:               21613   AIC:                         5.955e+05
Df Residuals:                   21601   BIC:                         5.956e+05
Df Model:                          11
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          6.418e+06   1.38e+05     46.627      0.000    6.15e+06    6.69e+06
bedrooms      -5.813e+04   2150.015    -27.038      0.000   -6.23e+04   -5.39e+04
bathrooms      6.615e+04   3737.401     17.701      0.000    5.88e+04    7.35e+04
sqft_lot          0.0371      0.055      0.671      0.502      -0.071       0.145
floors         5.498e+04   4020.903     13.673      0.000    4.71e+04    6.29e+04
waterfront     7.247e+05   1.86e+04     39.027      0.000    6.88e+05    7.61e+05
sqft_above      239.6824      3.895     61.538      0.000     232.048     247.317
sqft_basement   243.7353      4.812     50.654      0.000     234.304     253.167
yr_built      -3338.9292     71.492    -46.703      0.000   -3479.059   -3198.799
yr_renovated     11.9013      4.156      2.864      0.004       3.756      20.047
sqft_living15    90.4224      3.679     24.581      0.000      83.212      97.633
sqft_lot15       -0.7360      0.084     -8.731      0.000      -0.901      -0.571
==============================================================================
Omnibus:                    14160.528   Durbin-Watson:                   1.981
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           606691.177
Skew:                           2.579   Prob(JB):                         0.00
Kurtosis:                      28.438   Cond. No.                     4.40e+06
==============================================================================
```

Residual Analysis for Linear Regression

- Residual distribution

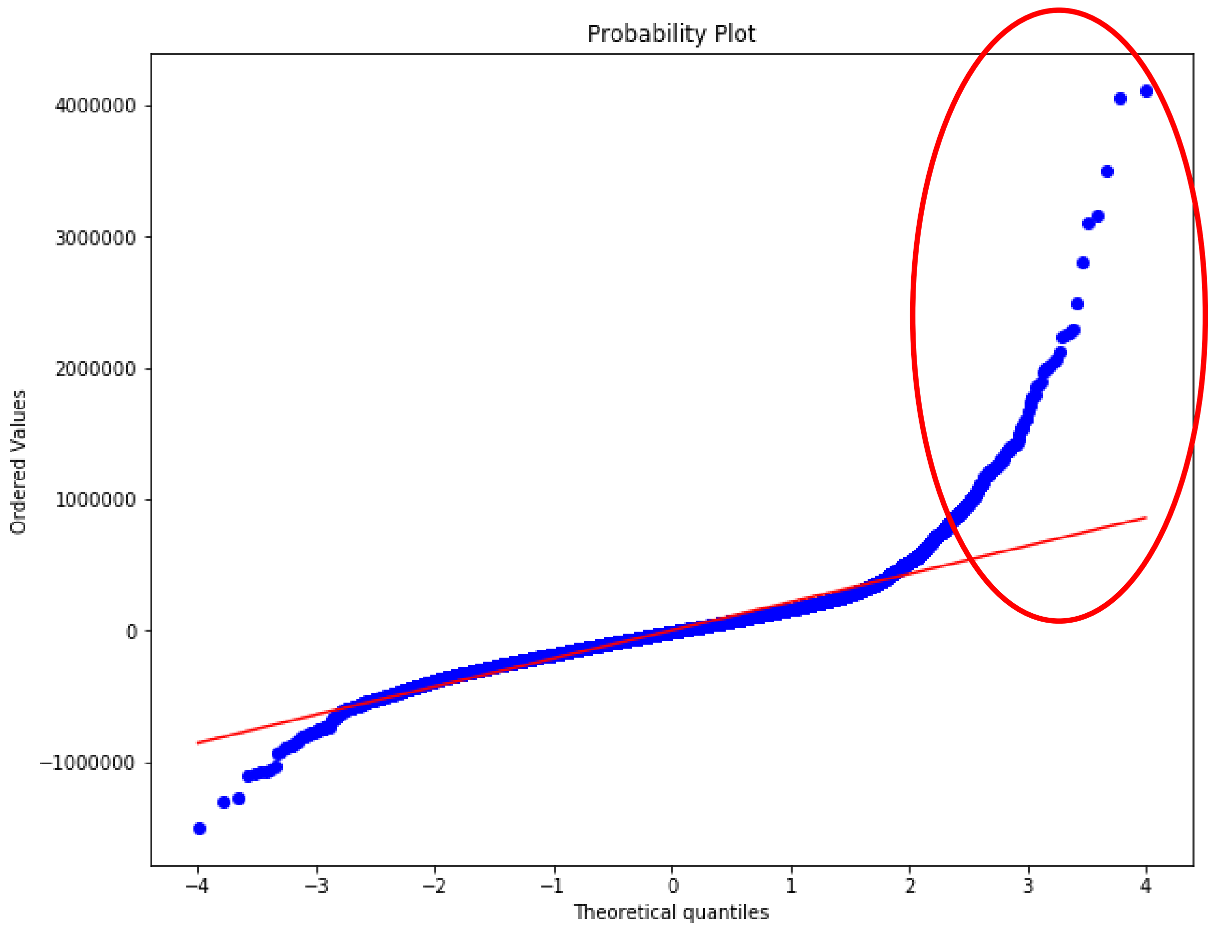


Residual Analysis for Linear Regression



Residual Analysis for Linear Regression

- Q-Q Plot

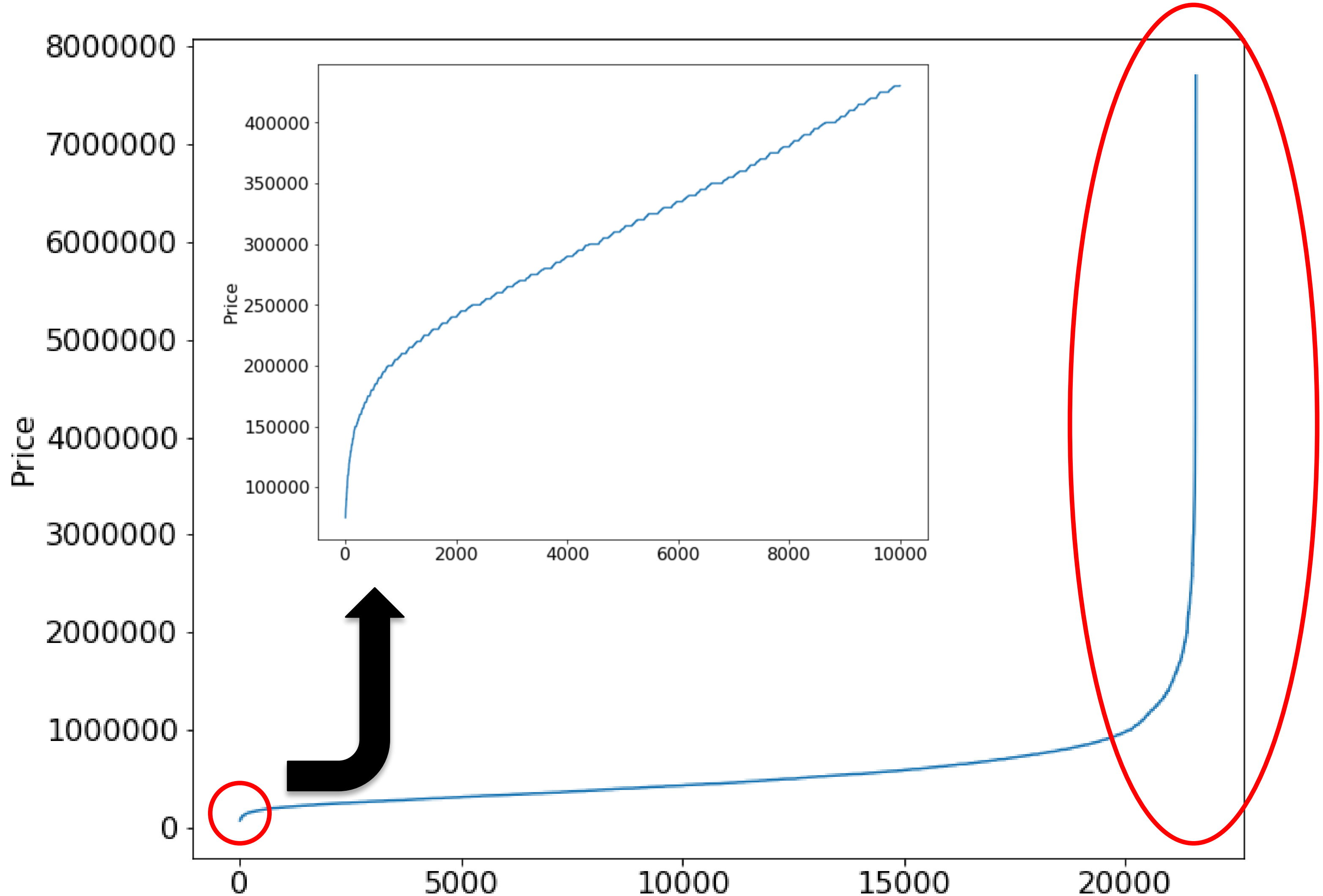


Residual Analysis for Linear Regression

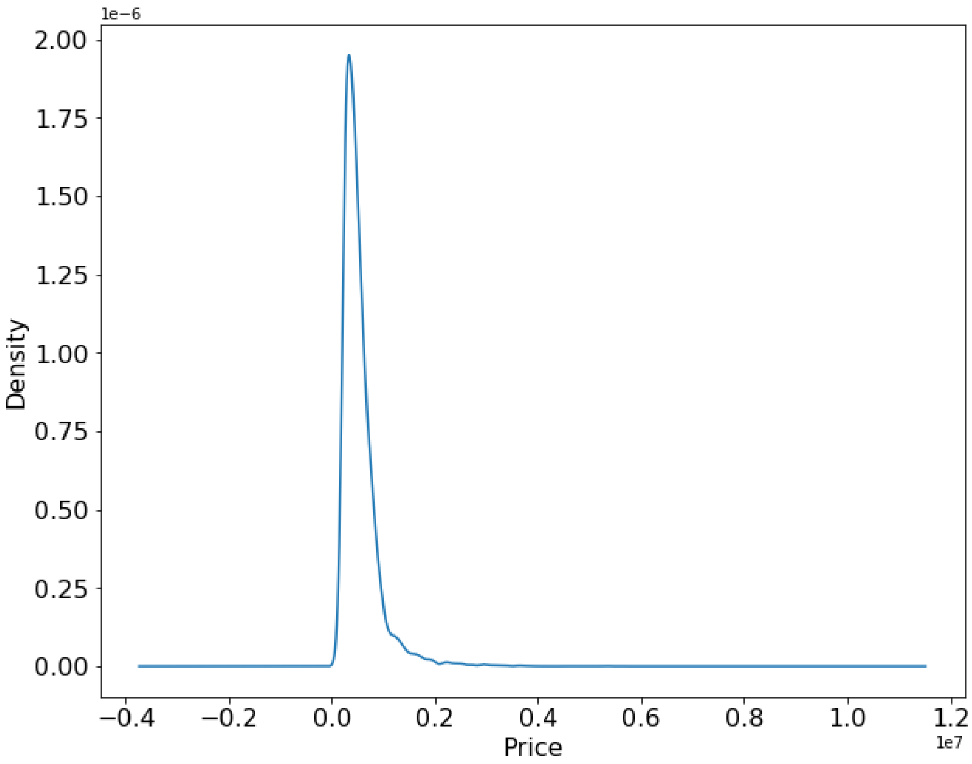
- Breusch-Pagan test using statsmodels
  - Use "het_breuschpagan"

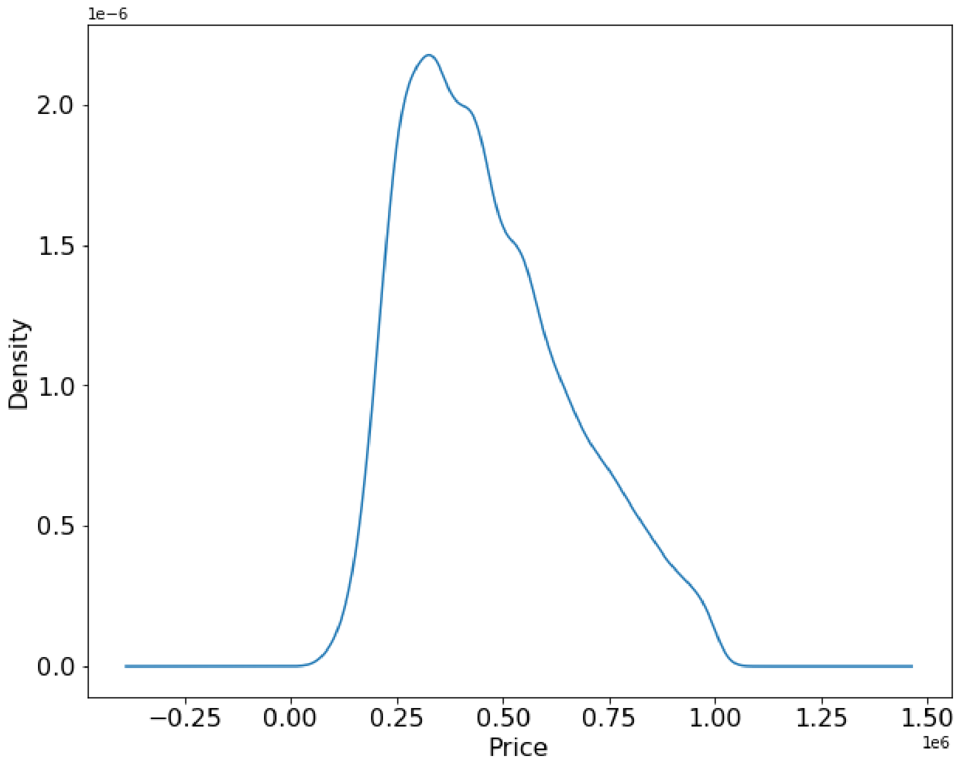
```
from statsmodels.stats import diagnostic
diagnostic.het_breuschpagan(resid, X)
```

Excluding Outliers



Excluding Outliers





Build a Model: Linear Regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.474
Model:                            OLS   Adj. R-squared:                  0.474
Method:                 Least Squares   F-statistic:                     1647.
Date:                Mon, 27 Jul 2020   Prob (F-statistic):               0.00
Time:                        11:38:29   Log-Likelihood:            -2.6723e+05
No. Observations:               20121   AIC:                         5.345e+05
Df Residuals:                   20109   BIC:                         5.346e+05
Df Model:                          11
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>¦t¦      [0.025      0.975]
---------------------------------------------------------------------------------
const          4.407e+06   9.05e+04     48.715      0.000    4.23e+06    4.58e+06
bedrooms      -2.231e+04   1399.096    -15.947      0.000   -2.51e+04   -1.96e+04
bathrooms      3.796e+04   2451.675     15.482      0.000    3.32e+04    4.28e+04
sqft_lot          0.1554      0.035      4.387      0.000       0.086       0.225
floors          7.09e+04   2594.964     27.320      0.000    6.58e+04     7.6e+04
waterfront     1.261e+05   1.87e+04      6.736      0.000    8.94e+04    1.63e+05
sqft_above       98.7145      2.823     34.972      0.000      93.182     104.247
sqft_basement   123.8863      3.364     36.826      0.000     117.292     130.480
yr_built      -2251.0211     47.115    -47.778      0.000   -2343.369   -2158.673
yr_renovated      2.7533      2.801      0.983      0.326      -2.737       8.244
sqft_living15    99.7699      2.582     38.641      0.000      94.709     104.831
sqft_lot15       -0.3191      0.054     -5.913      0.000      -0.425      -0.213
==============================================================================
Omnibus:                      437.916   Durbin-Watson:                   1.961
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              470.963
Skew:                           0.355   Prob(JB):                    5.39e-103
Kurtosis:                       3.238   Cond. No.                     4.41e+06
==============================================================================
```

Compare Models

Model 1

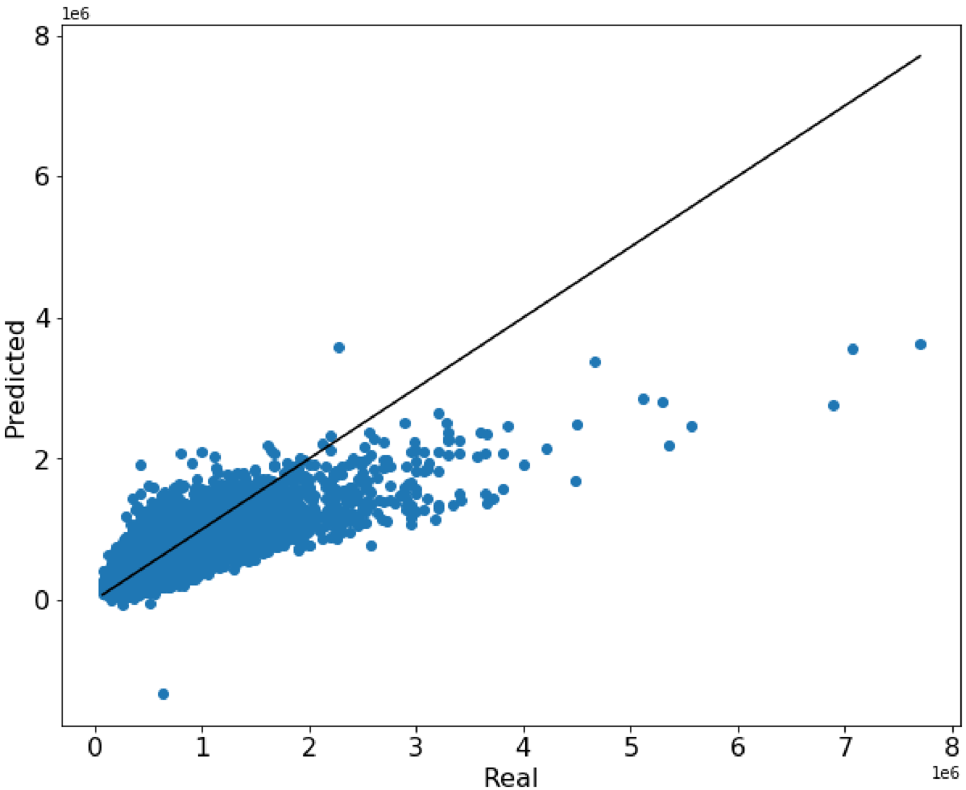
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.598
Model:                            OLS   Adj. R-squared:                  0.598
Method:                 Least Squares   F-statistic:                     2923.
Date:                Mon, 27 Jul 2020   Prob (F-statistic):               0.00
Time:                        12:16:59   Log-Likelihood:            -2.9775e+05
No. Observations:               21613   AIC:                         5.955e+05
Df Residuals:                   21601   BIC:                         5.956e+05
Df Model:                          11
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          6.418e+06   1.38e+05     46.627      0.000    6.15e+06    6.69e+06
bedrooms      -5.813e+04   2150.015    -27.038      0.000   -6.23e+04   -5.39e+04
bathrooms      6.615e+04   3737.401     17.701      0.000    5.88e+04    7.35e+04
sqft_lot          0.0371      0.055      0.671      0.502      -0.071       0.145
floors         5.498e+04   4020.903     13.673      0.000    4.71e+04    6.29e+04
waterfront     7.247e+05   1.86e+04     39.027      0.000    6.88e+05    7.61e+05
sqft_above      239.6824      3.895     61.538      0.000     232.048     247.317
sqft_basement   243.7353      4.812     50.654      0.000     234.304     253.167
yr_built      -3338.9292     71.492    -46.703      0.000   -3479.059   -3198.799
yr_renovated     11.9013      4.156      2.864      0.004       3.756      20.047
sqft_living15    90.4224      3.679     24.581      0.000      83.212      97.633
sqft_lot15       -0.7360      0.084     -8.731      0.000      -0.901      -0.571
==============================================================================
Omnibus:                    14160.528   Durbin-Watson:                   1.981
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           606691.177
Skew:                           2.579   Prob(JB):                         0.00
Kurtosis:                      28.438   Cond. No.                     4.40e+06
==============================================================================
```

Model 2

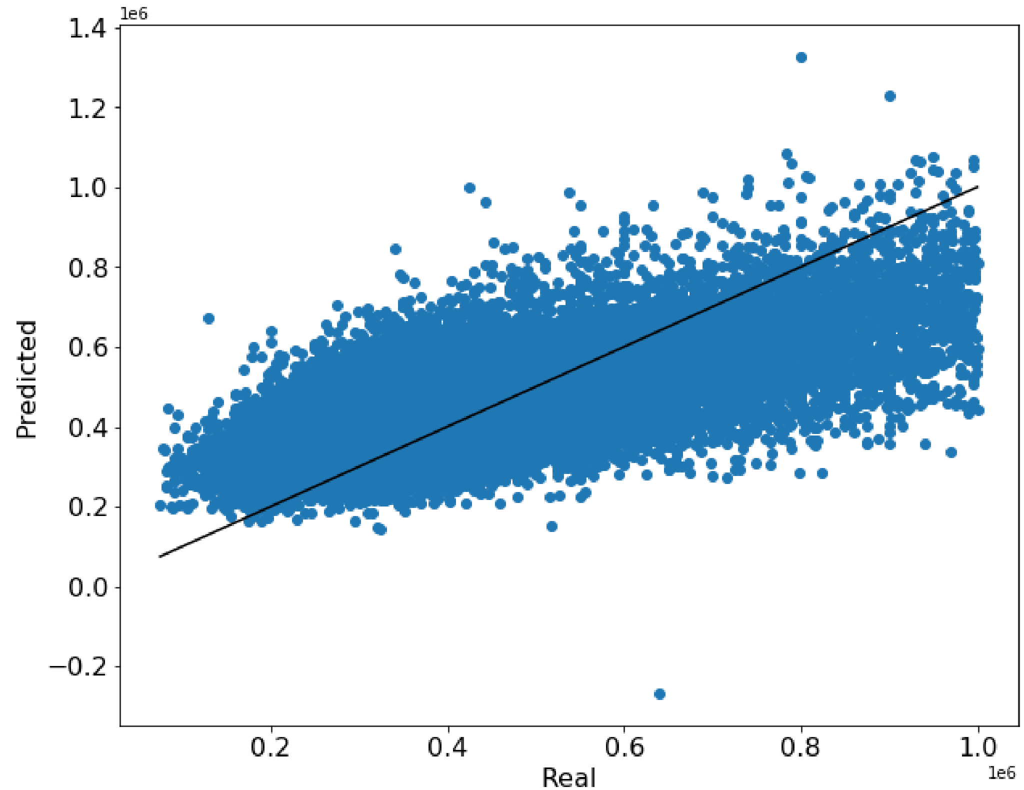
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.474
Model:                            OLS   Adj. R-squared:                  0.474
Method:                 Least Squares   F-statistic:                     1647.
Date:                Mon, 27 Jul 2020   Prob (F-statistic):               0.00
Time:                        11:38:29   Log-Likelihood:            -2.6723e+05
No. Observations:               20121   AIC:                         5.345e+05
Df Residuals:                   20109   BIC:                         5.346e+05
Df Model:                          11
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          4.407e+06   9.05e+04     48.715      0.000    4.23e+06    4.58e+06
bedrooms      -2.231e+04   1399.096    -15.947      0.000   -2.51e+04   -1.96e+04
bathrooms      3.796e+04   2451.675     15.482      0.000    3.32e+04    4.28e+04
sqft_lot          0.1554      0.035      4.387      0.000       0.086       0.225
floors          7.09e+04   2594.964     27.320      0.000    6.58e+04     7.6e+04
waterfront     1.261e+05   1.87e+04      6.736      0.000    8.94e+04    1.63e+05
sqft_above       98.7145      2.823     34.972      0.000      93.182     104.247
sqft_basement   123.8863      3.364     36.826      0.000     117.292     130.480
yr_built      -2251.0211     47.115    -47.778      0.000   -2343.369   -2158.673
yr_renovated      2.7533      2.801      0.983      0.326      -2.737       8.244
sqft_living15    99.7699      2.582     38.641      0.000      94.709     104.831
sqft_lot15       -0.3191      0.054     -5.913      0.000      -0.425      -0.213
==============================================================================
Omnibus:                      437.916   Durbin-Watson:                   1.961
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              470.963
Skew:                           0.355   Prob(JB):                    5.39e-103
Kurtosis:                       3.238   Cond. No.                     4.41e+06
==============================================================================
```

Compare Models

Model 1



Model 2



Question

- Why R^2 of the new model is lower than the previous model?

R^2

- Total variance: the total sum of squares

$$SST = \sum_i (y_i - \bar{y})^2$$

- Explained variance: the regression sum of squares, also called the explained sum of squares

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

- Residual variance: the sum of squares of residuals, also called the residual sum of squares

$$SSE = \sum_i (\hat{y}_i - y_i)^2$$

- R^2

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Model Evaluation: Regression

- Mean squared error
  - Risk metric corresponding to the expected value of the squared (quadratic) error loss or loss

$$MSE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Mean squared logarithmic error
  - Best to use when targets having exponential growth

$$MSLE = \frac{1}{n}\sum_{i=1}^{n}(\log_2(1 + y_i) - \log_e(1 + \hat{y}_i))^2$$

Model Evaluation: Regression

- Mean absolute error
  - Risk metric corresponding to the expected value of the absolute error loss or l_1-norm loss

$$MAE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

- Median absolute error
  - Calculated by taking the median of all absolute differences between the target and the prediction
  - Robust to outliers

$$MedAE(y, \hat{y}) = median(|y_0 - \hat{y}_0|, \dots, |y_n - \hat{y}_n|)$$

Assignment

Assignment 03

- Add categorical variables to variable set
  - ‘view’, ‘condition’, ‘grade’
  - Interpret the results

- Ideas to utilize zipcode, lat, and long
  - Without other resources
    - Data manipulation approach based on these three variables
  - With other resources
    - Additional useful information for these three variables

- Illustrate your ideas using Power Point
  - Some students have to create a video clip to explain their results and ideas