Introduction to Data Science

DSA1101

Semester 1, 2018/2019
Week 2

Multiple linear regression

Multiple linear regression

- Multiple linear regression is an analytical technique used to model the relationship between several input variables and a continuous outcome variable.
- A key assumption is that the relationships between the input variables and the outcome variable are linear.

Multiple linear regression: examples

Source: The Business Times

- **Real estate:** Linear regression analysis can be used to model residential home prices as a function of the home's living area.
- Such a model helps set or evaluate the list price of a home on the market.
- The model could be further improved by including other input variables such as number of bathrooms, number of bedrooms, lot size, school district rankings, crime statistics, and property taxes.

Multiple linear regression: examples

Source: The Straits Times

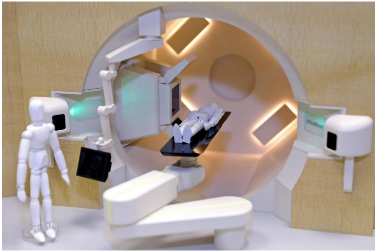
- **Demand forecasting:** Businesses and governments can use linear regression models to predict demand for goods and services.
- For example, coffee shops can appropriately prepare for the predicted type and quantity of food that customers will consume based upon the weather, the day of the week, whether an item is offered as a special, the time of day, and the reservation volume.

Multiple linear regression: examples

Source: The Straits Times

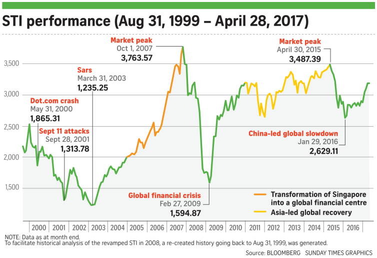
- Similar forecasting models can be built to predict taxi demand, emergency room visits, and ambulance dispatches.

Multiple linear regression: examples



Source: The Straits Times

- **Medical:** Linear regression model can be used to analyze the effect of a proposed radiation treatment on reducing tumor sizes.
- Multiple input variables might include duration of a single radiation treatment, frequency of radiation treatment, and patient attributes such as age or weight.



Source: Bloomberg, Sunday Times Graphics

- **Finance:** Multiple linear regression is used to model the relationships between stock market prices and other variables such as economic performance, interest rates and geopolitical risks.

Multiple linear regression: examples

Source: The Straits Times

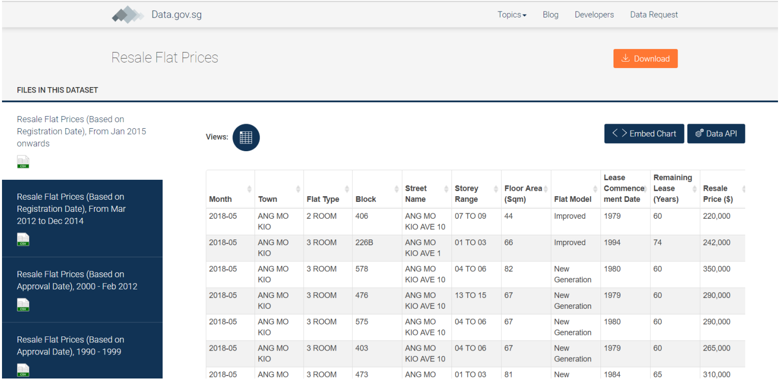
- **Pharmaceutical Industry:** Linear regression model can be used to analyze the clinical efficacies of drugs.
- Input variables may include age, gender and other patient characteristics such as blood pressure and blood sugar level.

Example closer to home...

Source: The Straits Times

- Data on resale HDB prices based on registration date is publicly available from `https://data.gov.sg/dataset/resale-flat-prices`.
- We have extracted a subset of all the resale records from March 2012 to December 2014 based on registration date.
- Available as the data set `HDBresale_reg.csv` on the course website.

Example closer to home...



Data.gov.sg

Topics Blog Developers Data Request

Resale Flat Prices

Download

FILES IN THIS DATASET

Resale Flat Prices (Based on Registration Date), From Jan 2015 onwards

Resale Flat Prices (Based on Registration Date), From Mar 2012 to Dec 2014

Resale Flat Prices (Based on Approval Date), 2000 - Feb 2012

Resale Flat Prices (Based on Approval Date), 1990 - 1999

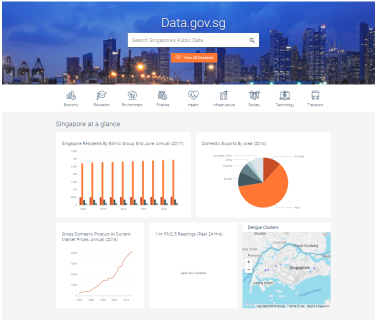
Views:

Embed Chart Data API

Month	Town	Flat Type	Block	Street Name	Storey Range	Floor Area (Sqm)	Flat Model	Lease Commencement Date	Remaining Lease (Years)	Resale Price ($)
2018-05	ANG MO KIO	2 ROOM	406	ANG MO KIO AVE 10	07 TO 09	44	Improved	1979	60	220,000
2018-05	ANG MO KIO	3 ROOM	226B	ANG MO KIO AVE 1	01 TO 03	66	Improved	1994	74	242,000
2018-05	ANG MO KIO	3 ROOM	578	ANG MO KIO AVE 10	04 TO 06	82	New Generation	1980	60	350,000
2018-05	ANG MO KIO	3 ROOM	476	ANG MO KIO AVE 10	13 TO 15	67	New Generation	1979	60	290,000
2018-05	ANG MO KIO	3 ROOM	575	ANG MO KIO AVE 10	04 TO 06	67	New Generation	1980	60	290,000
2018-05	ANG MO KIO	3 ROOM	403	ANG MO KIO AVE 10	04 TO 06	67	New Generation	1979	60	265,000
2018-05	ANG MO	3 ROOM	473	ANG MO	01 TO 03	81	New	1984	65	310,000

Source: data.gov.sg

GovTech datasets



Source: data.gov.sg

- https://data.gov.sg hosts many publicly available datasets for data analytics.

HDB resale data

- Let us take a closer look at the HDB resale dataset

```
>resale = read.csv("hdbresale_reg.csv")

> head(resale[,1:5])
    X   month          town flat_type block
1 580 2012-03 CENTRAL AREA    3 ROOM   640
2 581 2012-03 CENTRAL AREA    3 ROOM   640
3 582 2012-03 CENTRAL AREA    3 ROOM   668
4 583 2012-03 CENTRAL AREA    3 ROOM     5
5 584 2012-03 CENTRAL AREA    3 ROOM   271
6 585 2012-03 CENTRAL AREA    4 ROOM  671A
```

HDB resale data

- Let us take a closer look at the HDB resale dataset

```
> head(resale[,6:8])
     street_name storey_range floor_area_sqm
1      ROWELL RD     01 TO 05             74
2      ROWELL RD     06 TO 10             74
3     CHANDER RD     01 TO 05             73
4 TG PAGAR PLAZA     11 TO 15             59
5       QUEEN ST     11 TO 15             68
6     KLANG LANE     01 TO 05             75
```

HDB resale data

- Let us take a closer look at the HDB resale dataset

```
> head(resale[,9:11])
  flat_model lease_commence_date resale_price
1    Model A                1984       380000
2    Model A                1984       388000
3    Model A                1984       400000
4   Improved                1977       460000
5   Improved                1979       488000
6    Model A                2003       495000
```

Multiple linear regression

- Suppose we are interested to build a linear regression model that estimates a HDB unit's resale price as a function of `town`, `flattype` and `floorareainsquaremeters`.
- With more than one input variable, we will use *multiple linear regression*.

Multiple linear regression

- In the multiple linear regression model with p input variables,

$$y = \beta_0 + \beta_1 x(1) + \beta_2 x(2) + ... + \beta_p x(p) + \epsilon,$$

* y is the outcome variable
* $x(j)$ are the input variables, $j = 1, 2, ..., p$
* β_0 is the value of y when each $x(j)$ equals zero
* β_j is the change in y based on a unit change in $x(j)$ for $j = 1, 2, ..., p$
* ϵ is a random error term

Multiple linear regression

- For example, when there are three input variables, the linear model is

$$y = \beta_0 + \beta_1 x(1) + \beta_2 x(2) + \beta_3 x(3) + \epsilon.$$

- The parameters $(\beta_0, \beta_1, \beta_2, \beta_3)$ can be estimated by the method of least squares.
- We will review the least squares method in the simple linear regression case to consolidate understanding.

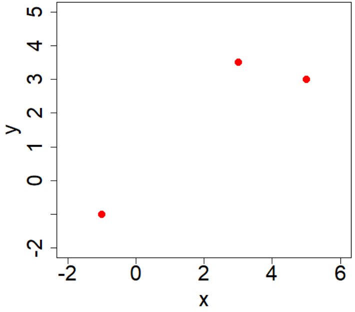
Review: least squares for simple linear regression

- Suppose we have three observations. Each observation has an outcome y and an input variable x.
- We are interested in the linear relationship

$$y_i \approx \beta_0 + \beta_1 x_i$$

- Since there is only one input variable, this is an example of simple linear model.

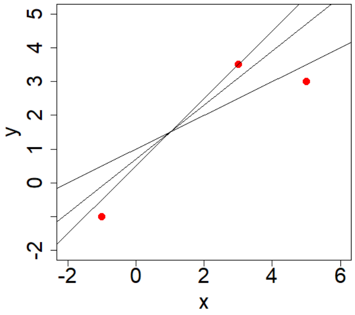
Review: least squares for simple linear regression



i	x_i	y_i
1	-1	-1
2	3	3.5
3	5	3

Plot of the three data points.

Review: least squares for simple linear regression

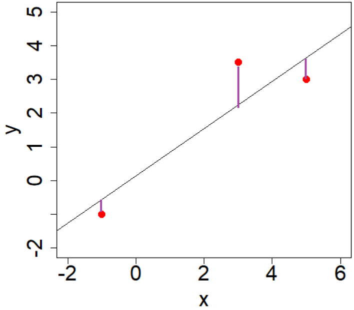


- We are interested in the linear relationship

$$y_i \approx f(x_i) = \beta_0 + \beta_1 x_i$$

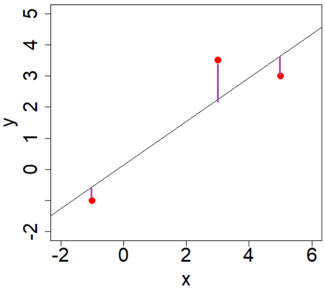
- Recall that the above is actually the formula for a straight line.
- There are many different lines that can be used to model x and y, as shown in the plot.

Review: least squares for simple linear regression



- Intuitively, we want the line to be as close to the data points as possible.
- This "closeness" can be measured in terms of the vertical distance between each point to the line (represented by the length of the purple lines).

Review: least squares for simple linear regression



			fitted values	difference in y
i	x_i	y_i	$\beta_0 + \beta_1 x_i$	residual: $e_i = y_i - (\beta_0 + \beta_1 x_i)$
1	-1	-1	$\beta_0 + (-1)\beta_1$	$-1 - (\beta_0 + (-1)\beta_1) = -1 - \beta_0 + \beta_1$
2	3	3.5	$\beta_0 + (3)\beta_1$	$3.5 - (\beta_0 + (3)\beta_1) = 3.5 - \beta_0 - 3\beta_1$
3	5	3	$\beta_0 + (5)\beta_1$	$3 - (\beta_0 + (5)\beta_1) = 3 - \beta_0 - 5\beta_1$

Review: least squares for simple linear regression

- Now the residual for each point may be positive or negative:
- We do not want the residuals to "cancel off" each other, so we square each of them, leading to the squared residuals.

i	residual: $e_i = y_i - (\beta_0 + \beta_1 x_i)$	squared residual: e_i^2
1	$-1 - (\beta_0 + (-1)\beta_1) = -1 - \beta_0 + \beta_1$	$[-1 - \beta_0 + \beta_1]^2$
2	$3.5 - (\beta_0 + (3)\beta_1) = 3.5 - \beta_0 - 3\beta_1$	$[3.5 - \beta_0 - 3\beta_1]^2$
3	$3 - (\beta_0 + (5)\beta_1) = 3 - \beta_0 - 5\beta_1$	$[3 - \beta_0 - 5\beta_1]^2$

Review: least squares for simple linear regression

i	residual: $e_i = y_i - (\beta_0 + \beta_1 x_i)$	squared residual: e_i^2
1	$-1 - (\beta_0 + (-1)\beta_1) = -1 - \beta_0 + \beta_1$	$[-1 - \beta_0 + \beta_1]^2$
2	$3.5 - (\beta_0 + (3)\beta_1) = 3.5 - \beta_0 - 3\beta_1$	$[3.5 - \beta_0 - 3\beta_1]^2$
3	$3 - (\beta_0 + (5)\beta_1) = 3 - \beta_0 - 5\beta_1$	$[3 - \beta_0 - 5\beta_1]^2$

- To express the total magnitude of the deviations, we sum up the squared residuals for all the data points.
- The resulting sum is the Residual Sum of Squares, abbreviated as RSS
- In the above example,

$$RSS = e_1^2 + e_1^2 + e_3^2$$

$$= [-1 - \beta_0 + \beta_1]^2 + [3.5 - \beta_0 - 3\beta_1]^2 + [3 - \beta_0 - 5\beta_1]^2 \,.$$

Review: least squares for simple linear regression

- We now seek the values of β_0 and β_1 such that the RSS, given by

  $$RSS = [-1 - \beta_0 + \beta_1]^2 + [3.5 - \beta_0 - 3\beta_1]^2 + [3 - \beta_0 - 5\beta_1]^2 \,,$$

  is minimized.
- This process is known as the method of least squares.

Review: least squares for simple linear regression

- We now have a function in terms of β_0 and β_1. Let's call it $h(\beta_0, \beta_1)$ so that

$$h(\beta_0, \beta_1) = [-1 - \beta_0 + \beta_1]^2 + [3.5 - \beta_0 - 3\beta_1]^2 + [3 - \beta_0 - 5\beta_1]^2 .$$

- To find the minimum value of $h(\beta_0, \beta_1)$, first differentiate with respect to β_0, while holding β_1 constant:

$$\begin{aligned}\frac{\partial h(\beta_0, \beta_1)}{\partial \beta_0} &= 2\,[-1 - \beta_0 + \beta_1]\,(-1) \\ &\quad + 2\,[3.5 - \beta_0 - 3\beta_1]\,(-1) + 2\,[3 - \beta_0 - 5\beta_1]\,(-1) \\ &= 2 + 2\beta_0 - 2\beta_1 - 7 + 2\beta_0 + 6\beta_1 - 6 + 2\beta_0 + 10\beta_1 \\ &= -11 + 6\beta_0 + 14\beta_1.\end{aligned}$$

Review: least squares for simple linear regression

- Then differentiate with respect to β_1, while holding β_0 constant:

$$
\begin{aligned}
\frac{\partial h(\beta_0, \beta_1)}{\partial \beta_1} &= 2\left[-1 - \beta_0 + \beta_1\right](1) \\
&+2\left[3.5 - \beta_0 - 3\beta_1\right](-3) + 2\left[3 - \beta_0 - 5\beta_1\right](-5) \\
&= -2 - 2\beta_0 + 2\beta_1 - 21 + 6\beta_0 + 18\beta_1 - 30 + 10\beta_0 + 50\beta_1 \\
&= -53 + 14\beta_0 + 70\beta_1.
\end{aligned}
$$

Review: least squares for simple linear regression

- Finally, by setting both the derivative to zero, we have the system of equations

$$
\begin{aligned}
-11 + 6\beta_0 + 14\beta_1 &= 0 \\
-53 + 14\beta_0 + 70\beta_1 &= 0
\end{aligned}
$$

- Solving the equations,

$$
\begin{aligned}
\beta_0 &= \frac{11}{6} - \frac{14}{6}\beta_1 \\
-53 + 14\beta_0 + 70\beta_1 &= 0,
\end{aligned}
$$

- leads to the *least squares* estimates

$$
\begin{aligned}
\beta_0 &\approx 0.1250 \\
\beta_1 &\approx 0.7321
\end{aligned}
$$

Review: least squares for simple linear regression

- We usually add the "hat" sign on top of parameter to denote estimated values, so the least squares estimates are denoted as

$$
\begin{aligned}
\hat{\beta}_0 &\approx 0.1250 \\
\hat{\beta}_1 &\approx 0.7321.
\end{aligned}
$$

Review: least squares for simple linear regression

- We can check that the least squares estimates we computed are equivalent to those returned by `lm()` function in R:

```
> x<-c( -1, 3, 5)
> y<-c( -1, 3.5, 3)
> lm(y~x);

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     0.1250       0.7321
```

Review: least squares for simple linear regression

- Notice that on slide 22, we begin with this table:

i	x_i	y_i	$\beta_0 + \beta_1 x_i$	residual: $e_i = y_i - (\beta_0 + \beta_1 x_i)$
1	-1	-1	$\beta_0 + (-1)\beta_1$	$-1 - \beta_0 + \beta_1$
2	3	3.5	$\beta_0 + (3)\beta_1$	$3.5 - \beta_0 - 3\beta_1$
3	5	3	$\beta_0 + (5)\beta_1$	$3 - \beta_0 - 5\beta_1$

- However, the values for β_0 and β_1 were unknown.
- Now that we have obtained the least squares estimates $\hat{\beta}_0 \approx 0.1250$ and $\hat{\beta}_1 \approx 0.7321$, we can plug those values into the table above!

Review: least squares for simple linear regression

i	x_i	y_i	$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$	residual: $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$
1	-1	-1	-0.6071	-0.3929
2	3	3.5	2.3213	1.1787
3	5	3	3.7855	-0.7855

- The column for $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ contains the fitted values for outcome y.
- The column for $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ contains the residuals after fitting the simple linear model.

Review: least squares for simple linear regression

i	x_i	y_i	$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$	residual: $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$
1	-1	-1	-0.6071	-0.3929
2	3	3.5	2.3213	1.1787
3	5	3	3.7855	-0.7855

- We see that R can also output the fitted values and residuals directly after fitting the linear model:

```
> x<-c( -1, 3, 5)
> y<-c( -1, 3.5, 3)
> lmout <-lm(y~x)
> lmout$fitted.values
         1          2          3
-0.6071429  2.3214286  3.7857143
> lmout$residuals
         1          2          3
-0.3928571  1.1785714 -0.7857143
```

Review: least squares for simple linear regression

- We can follow up by plotting the residuals against the fitted outcome values for model diagnostics, as discussed in the previous lecture.

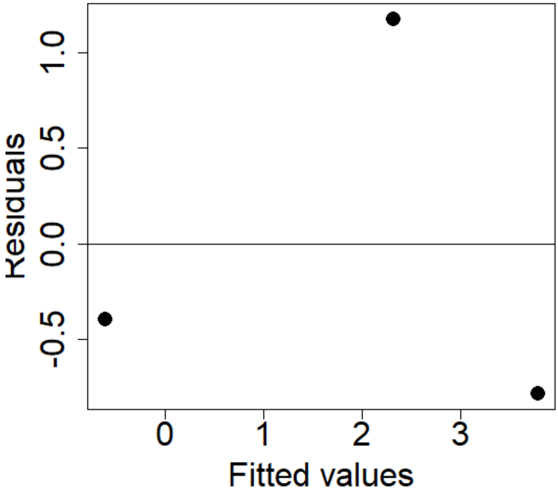
```
x<-c( -1, 3, 5)
y<-c( -1, 3.5, 3)

lmout <-lm(y~x)

plot(x=lmout$fitted.values, y=lmout$residuals,
      xlab="Fitted values", ylab="Residuals",
      cex=2, cex.lab=2, cex.axis=2, pch=16)

abline(0,0)
```

Review: least squares for simple linear regression



Back to multiple linear regression...

- In the multiple linear regression model with p input variables,

$$y = \beta_0 + \beta_1 x(1) + \beta_2 x(2) + ... + \beta_p x(p) + \epsilon,$$

∗ y is the outcome variable

∗ $x(j)$ are the input variables, $j = 1, 2, ..., p$

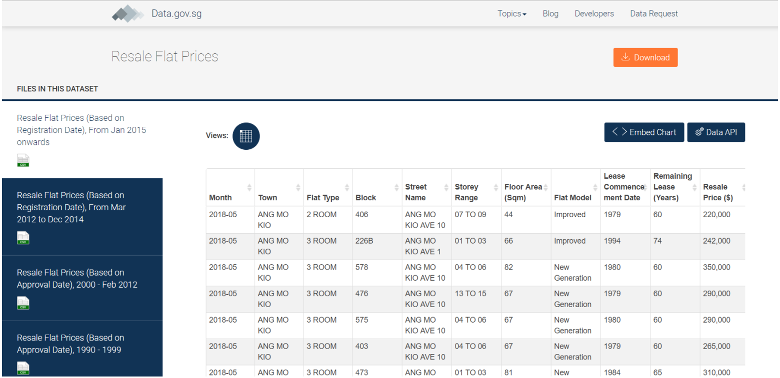
∗ β_0 is the value of y when each $x(j)$ equals zero

∗ β_j is the change in y based on a unit change in $x(j)$ for $j = 1, 2, ..., p$

∗ ϵ is a random error term

- We can also estimate the unknown parameters in the multiple linear regression model via the method of least squares

Back to our HDB resale data example...



Data.gov.sg

Topics Blog Developers Data Request

Resale Flat Prices

Download

FILES IN THIS DATASET

Resale Flat Prices (Based on Registration Date), From Jan 2015 onwards

Resale Flat Prices (Based on Registration Date), From Mar 2012 to Dec 2014

Resale Flat Prices (Based on Approval Date), 2000 - Feb 2012

Resale Flat Prices (Based on Approval Date), 1990 - 1999

Views:

Embed Chart

Data API

Month	Town	Flat Type	Block	Street Name	Storey Range	Floor Area (Sqm)	Flat Model	Lease Commencement Date	Remaining Lease (Years)	Resale Price ($)
2018-05	ANG MO KIO	2 ROOM	406	ANG MO KIO AVE 10	07 TO 09	44	Improved	1979	60	220,000
2018-05	ANG MO KIO	3 ROOM	226B	ANG MO KIO AVE 1	01 TO 03	66	Improved	1994	74	242,000
2018-05	ANG MO KIO	3 ROOM	578	ANG MO KIO AVE 10	04 TO 06	82	New Generation	1980	60	350,000
2018-05	ANG MO KIO	3 ROOM	476	ANG MO KIO AVE 10	13 TO 15	67	New Generation	1979	60	290,000
2018-05	ANG MO KIO	3 ROOM	575	ANG MO KIO AVE 10	04 TO 06	67	New Generation	1980	60	290,000
2018-05	ANG MO KIO	3 ROOM	403	ANG MO KIO AVE 10	04 TO 06	67	New Generation	1979	60	265,000
2018-05	ANG MO	3 ROOM	473	ANG MO	01 TO 03	81	New	1984	65	310,000

Source: data.gov.sg

Multiple linear regression

- Suppose we are interested to build a linear regression model that estimates a HDB unit's resale price as a function of `town`, `flattype` and `floorareainsquaremeters`.
- With more than one input variable, we will use *multiple linear regression*.

Multiple linear regression

```
> resale = read.csv("hdbresale_reg.csv")
> head(resale[,c(3,4,8,11)])
          town flat_type floor_area_sqm resale_price
1 CENTRAL AREA    3 ROOM             74       380000
2 CENTRAL AREA    3 ROOM             74       388000
3 CENTRAL AREA    3 ROOM             73       400000
4 CENTRAL AREA    3 ROOM             59       460000
5 CENTRAL AREA    3 ROOM             68       488000
6 CENTRAL AREA    4 ROOM             75       495000
```

HDB resale data

- Note that in our `resale` dataset, variables such as `town` and `flat_type` are called *categorical variables*, since they consist of different categories instead of numerical values.
- The function `levels()` in R will display all the different categories in a variable.

```
> levels(resale$town)
[1] "CENTRAL AREA" "JURONG EAST"  "WOODLANDS"
> levels(resale$flat_type)
[1] "2 ROOM"     "3 ROOM"     "4 ROOM"     "5 ROOM"     "
    EXECUTIVE"
```

HDB resale data

- In R, we can perform multiple linear regression using the `lm()` function:

```
> lm(resale_price~town+floor_area_sqm+flat_type, data=resale)

Call:
lm(formula = resale_price ~ town + floor_area_sqm + flat_type,
    data = resale)

Coefficients:
      (Intercept)      townJURONG EAST
           193438              -122748
    townWOODLANDS       floor_area_sqm
          -169896                 2526
  flat_type3 ROOM      flat_type4 ROOM
            98827               129929
  flat_type5 ROOM  flat_typeEXECUTIVE
           142570               214622
```

Categorical variables

- It is not meaningful to assign numerical values to a categorical variable such as `town`. For example, it is not meaningful to consider Jurong East to be one unit greater than Woodlands and two units greater than Central Area
- HDB resales that took place in the Central Areas will be regarded as the reference case.
- So for example, the coefficient estimate of -122748 for `townJURONGEAST` means that HDB resale price in Jurong East is on average \$122748 less than the resale price in Central areas, other variables being held constant.
- Similarly, the coefficient estimate of -169896 for `townWOODLANDS` means that HDB resale price in Woodlands is on average \$169896 less than the resale price in Central areas, other variables being held constant.

Multiple linear regression

- Suppose we are interested in computing *confidence intervals* for the parameter estimates from our multiple linear regression model
- Similar to the simple linear regression setting, we assume that the error terms are independent and normally distributed with mean zero and constant variances.

Multiple linear regression

- R simplifies the computation of confidence intervals on the parameters with the use of the `confint()` function.
- For example, the following R command provides 95% confidence intervals on our parameter estimates:

```
> out = lm(resale_price~town+floor_area_sqm+flat_type,
    data=resale)
> confint(out, level= .95)
                          2.5 %      97.5 %
(Intercept)          180791.622  206084.668
townJURONG EAST     -127894.901 -117601.145
townWOODLANDS       -174833.773 -164959.166
floor_area_sqm         2403.141    2649.163
flat_type3 ROOM       87053.089  110601.512
flat_type4 ROOM      117163.666  142693.511
flat_type5 ROOM      128354.088  156786.844
flat_typeEXECUTIVE   197641.718  231602.642
```

Multiple linear regression

- R simplifies the computation of confidence intervals on the parameters with the use of the `confint()` function.
- For example, the following R command provides 95% confidence intervals on our parameter estimates:

```
> out = lm(resale_price~town+floor_area_sqm+flat_type,
    data=resale)
> confint(out, level= .95)
                          2.5 %      97.5 %
(Intercept)          180791.622  206084.668
townJURONG EAST     -127894.901 -117601.145
townWOODLANDS       -174833.773 -164959.166
floor_area_sqm         2403.141    2649.163
flat_type3 ROOM       87053.089  110601.512
flat_type4 ROOM      117163.666  142693.511
flat_type5 ROOM      128354.088  156786.844
flat_typeEXECUTIVE   197641.718  231602.642
```

Multiple linear regression

- R also computes the *p-value* for testing $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ for $j = 0, 1, 2, \ldots, p$.

```
> out = lm(resale_price~town+floor_area_sqm+flat_type, data=resale)
> summary(out)

Call:
lm(formula = resale_price ~ town + floor_area_sqm + flat_type,
    data = resale)

Residuals:
    Min      1Q  Median      3Q     Max
-139026  -23350   -1453   19284  336649

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         193438.15    6451.13   29.98   <2e-16 ***
townJURONG EAST    -122748.02    2625.48  -46.75   <2e-16 ***
townWOODLANDS      -169896.47    2518.57  -67.46   <2e-16 ***
floor_area_sqm        2526.15      62.75   40.26   <2e-16 ***
flat_type3 ROOM      98827.30    6006.16   16.45   <2e-16 ***
flat_type4 ROOM     129928.59    6511.53   19.95   <2e-16 ***
flat_type5 ROOM     142570.47    7251.94   19.66   <2e-16 ***
flat_typeEXECUTIVE  214622.18    8661.93   24.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Confidence interval on the expected outcome

- Suppose we are interested in the expected resale price for 3-room HDB units in the Central Area with floor area of 70 square meters.
- The `predict()` function in R can provide a 95% confidence interval on this expected resale price

```
> out = lm(resale_price~town+floor_area_sqm+flat_type,
    data=resale)
>
> town="CENTRAL AREA"
> flat_type="3 ROOM"
> floor_area_sqm=70
>
> new_pt <- data.frame(town,flat_type,floor_area_sqm)
> conf_int_pt <- predict(out,new_pt,level=.95,interval=
    "confidence")
>
> conf_int_pt
       fit    lwr      upr
1 469096.1 464369 473823.2
```

Confidence interval on a particular outcome

- Suppose we are interested in the resale price for a particular 3-room HDB unit in the Central Area with floor area of 70 square meters.
- The `predict()` function in R can also provide a 95% prediction interval on this resale price

```
> out = lm(resale_price~town+floor_area_sqm+flat_type, data=resale)
>
> town="CENTRAL AREA"
> flat_type="3 ROOM"
> floor_area_sqm=70
>
> new_pt <- data.frame(town,flat_type,floor_area_sqm)
> conf_int_pt <- predict(out,new_pt,level=.95,interval="prediction")
>
> conf_int_pt
       fit      lwr      upr
1 469096.1 391692.2 546499.9
```

Model Diagnostics: Evaluating the Residuals

- We have made assumptions on the error terms in the multiple linear regression model
- We can plot the residuals against fitted values for visual inspection

```
> out = lm(resale_price~town+floor_area_sqm+flat_type, data=resale)
>
> plot(x=out$fitted.values, y=out$residuals,
+       xlab="Fitted values", ylab="Residuals", col="red")
> abline(0,0)
```

Model Diagnostics: Evaluating the Residuals