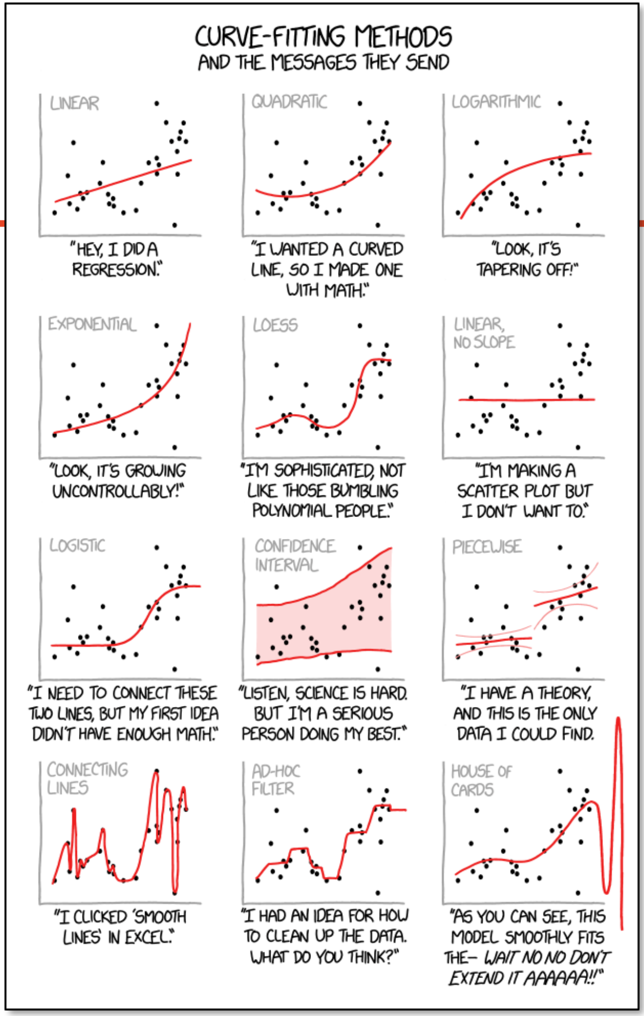
STAT5003 - Week 1

Line of best fit

Regression

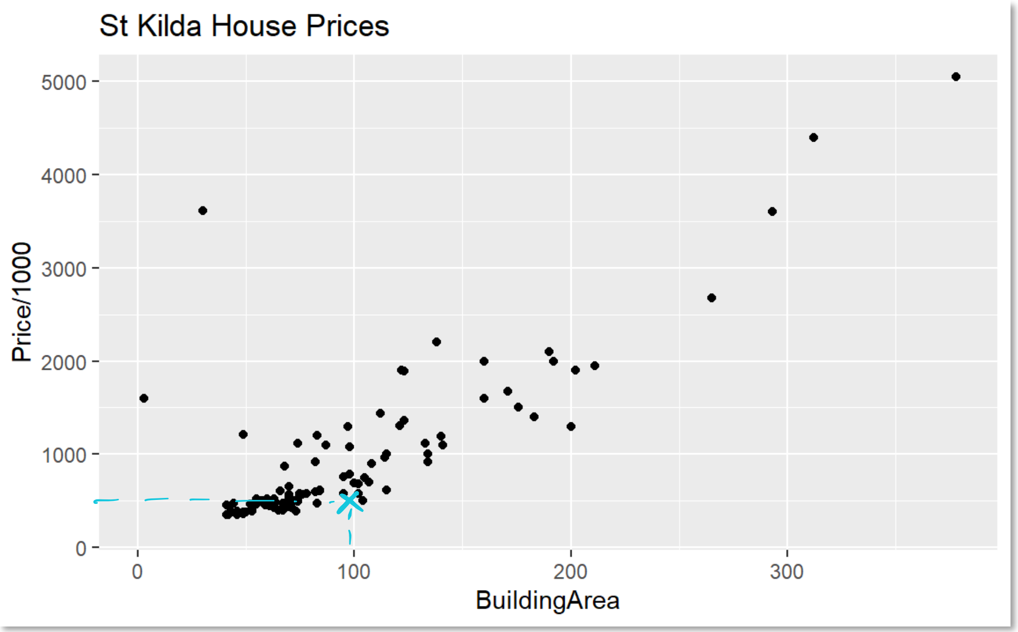
- Numerically fitting the model is easy
- Knowing how to appropriately fit the model is where you add value.



Source: https://xkcd.com/2048/

The prediction problem

- What is the price of a 100 sqm house in St Kilda?



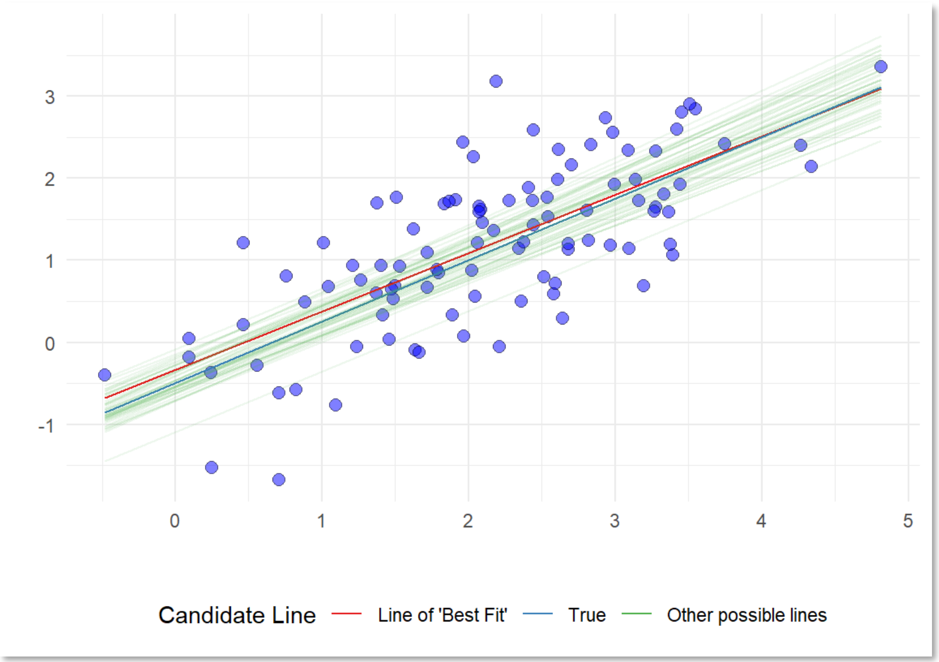
The linear regression model

seen it a thousand times...

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

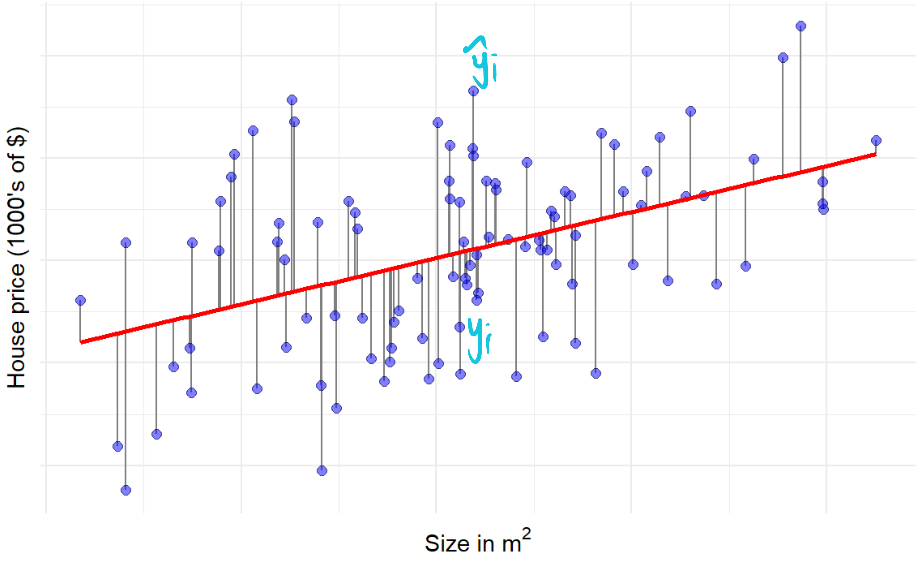
- X is the **predictor** (feature or independent variable)
- Y is the **response** (target or dependent variable)
- β_0 is the **intercept** of the regression line
- β_1 is the **slope** of the regression line
- ε is the **unexplained variation** or random error.

Performance of regression estimates



- Data was simulated from model $Y = -0.5 + 0.75X + \varepsilon$
- True line shown in blue
- Standard linear regression fit shown in red
- Why not one of the green lines?

Need an optimal criterion



- Easiest mathematical solution is the **least squares criterion**
  - Minimise the residual sum of squares

$$\boxed{RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2}$$

UNSW

Recall MATH2831 content for derivation of RSS

$e_i^2 = y_i - \hat{y}_i$

Least squares equations

- Can show by simple calculus the following:
  - Regression (slope) coefficient: $\boxed{b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{cov(x,y)}{var(x)}}$
  - Intercept: $\boxed{b_0 = \overline{y} - b_1\overline{x}}$
- This leads to the estimated regression line:

$$\hat{y} = b_0 + b_1 x$$

- Least squares regression line since it **minimises** the residual sum of squares.

THE UNIVERSITY OF
SYDNEY

Simple linear regression

St Kilda house price data

```
head(st.kilda.data)
  BuildingArea   Price Type
1          112 1440000    h
2           70  540000    u
3           70  650000    u
4           49 1210000    u
5          183 1400000    t
6           71  430000    u
```

Fitting a linear model

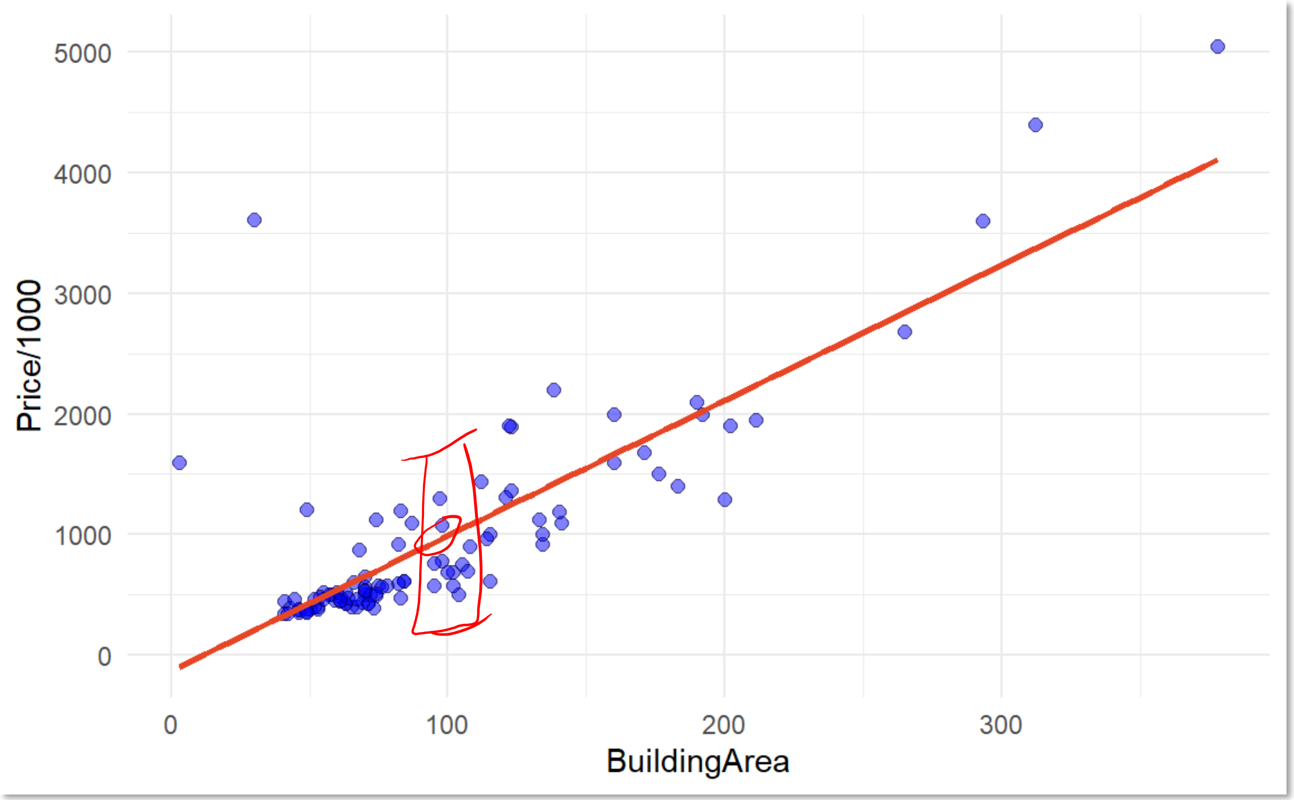
```r
lm.fit <- lm(Price ~ BuildingArea, data = st.kilda.data)
summary(lm.fit)
```

```
Call:
lm(formula = Price ~ BuildingArea, data = st.kilda.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-817415 -201614  -85181   19895 3403199 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -129484.0    91775.9  -1.411    0.161    
BuildingArea  11209.5      799.8  14.015   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 490300 on 99 degrees of freedom
Multiple R-squared:  0.6649,	Adjusted R-squared:  0.6615 
F-statistic: 196.4 on 1 and 99 DF,  p-value: < 2.2e-16
```

$$\hat{y} = -129484 + 11209 \times \text{BuildingArea}$$

$$\widehat{\text{Price}} = -129484 + 11209 \times \text{Area}$$

Linear regression fit



Standard error of population mean

- Consider single population estimation problem .
  - Wish to estimate some mean, μ, of some random variable Y.
  - If Y_i is sampled then $\hat{\mu} = \overline{Y}$ estimates μ with
  - $\boxed{Var(\hat{\mu}) = \left(SE(\hat{\mu})\right)^2 = \frac{\sigma^2}{n}}$
  - σ^2 is the variance of Y_i
  - n is the sample size.

$SE(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}$.

Standard error of regression coefficient estimates

- Same concept applies to the regression estimates

$$\boxed{SE(\widehat{\beta_0}) = \sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}}$$

$$\boxed{SE(\widehat{\beta_1}) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}}$$

where $\sigma^2 = Var(\varepsilon)$

$Var(x_i)\uparrow \quad se(x_i)\downarrow$

- As $n \to \infty$, $SE(\hat{\beta}_0) \to 0$ and $SE(\hat{\beta}_1) \to 0$
- Interestingly, if the x_i are more spread out, the standard errors will be smaller
  - more leverage to estimate the parameters.

Using standard errors to compute confidence intervals

```
summary(lm.fit) # Truncated output with coefficient table
...
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -129484.0    91775.9  -1.411    0.161
BuildingArea   11209.5      799.8  14.015   <2e-16 ***
---

Residual standard error: 490300 on 99 degrees of freedom
...
```

confidence intervals for the slope here.

$n-2$.

- We can use the standard error to estimate the 95% confidence interval as:
  - $\left(\hat{\beta}_1 - t_{n-2,0.975} SE(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2,0.975} SE(\hat{\beta}_1)\right) = b_1 \pm t_{n-2,0.975} SE(b_1) = b_1 \pm t_{99,0.975} SE(b_1)$
- In our housing example, the 95% confidence interval for the coefficient of `BuildingArea` is [9622.6968, 12796.3032]

Confidence intervals of regression coefficients

- More directly in `R` code, use the `confint` function.

```
confint(lm.fit)
```

```
                  2.5 %   97.5 %
(Intercept)  -311587.233 52619.18
BuildingArea    9622.491 12796.50
```

- Is exact, no precision lost to rounding error and easy to change

```
confint(lm.fit, level = 0.99)
```

```
                 0.5 %    99.5 %
(Intercept)  -370524.63 111556.57
BuildingArea    9108.86  13310.13
```

Is `BuildingArea` a good predictor of price?

```
summary(lm.fit) # truncated for coefficient table
...
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -129484.0    91775.9  -1.411    0.161
BuildingArea   11209.5      799.8  14.015   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
...
```

- Linear regression assumes $Y = \beta_0 + \beta_1 X + \varepsilon$
- If `BuildingArea` is not linearly related to `Price`, then $\beta_1 = 0$.
- Can conduct a test of significance $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$
- Can conduct a hypothesis test by computing the t-statistic:
- $\boxed{t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \overset{H_0}{=} \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}}$

Is `BuildingArea` a good predictor of price?

```
summary(lm.fit) # truncated for coefficient table
...
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -129484.0    91775.9  -1.411    0.161
BuildingArea   11209.5      799.8  14.015   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
...
```

p-values measure the probability of the t-statistic.

- The **p-values** for each significance test in the last column.
- Recall, **p-value** gives the probability of observing your test statistic (and other scenarios support H_1) assuming H_0 is true.
- Small p-value here gives very little evidence to support the claim that there is no relationship between `Price` and `BuildingArea`

Estimating the price of a 100 sqm house in St Kilda

```
new.100 <- data.frame(BuildingArea = 100)
predict(lm.fit, new.100, interval = "confidence")
       fit      lwr     upr
1 991465.5 894562.7 1088368
predict(lm.fit, new.100, interval = "prediction")
       fit      lwr     upr
1 991465.5 13820.26 1969111
```

this means one of the predictors is set to 100.

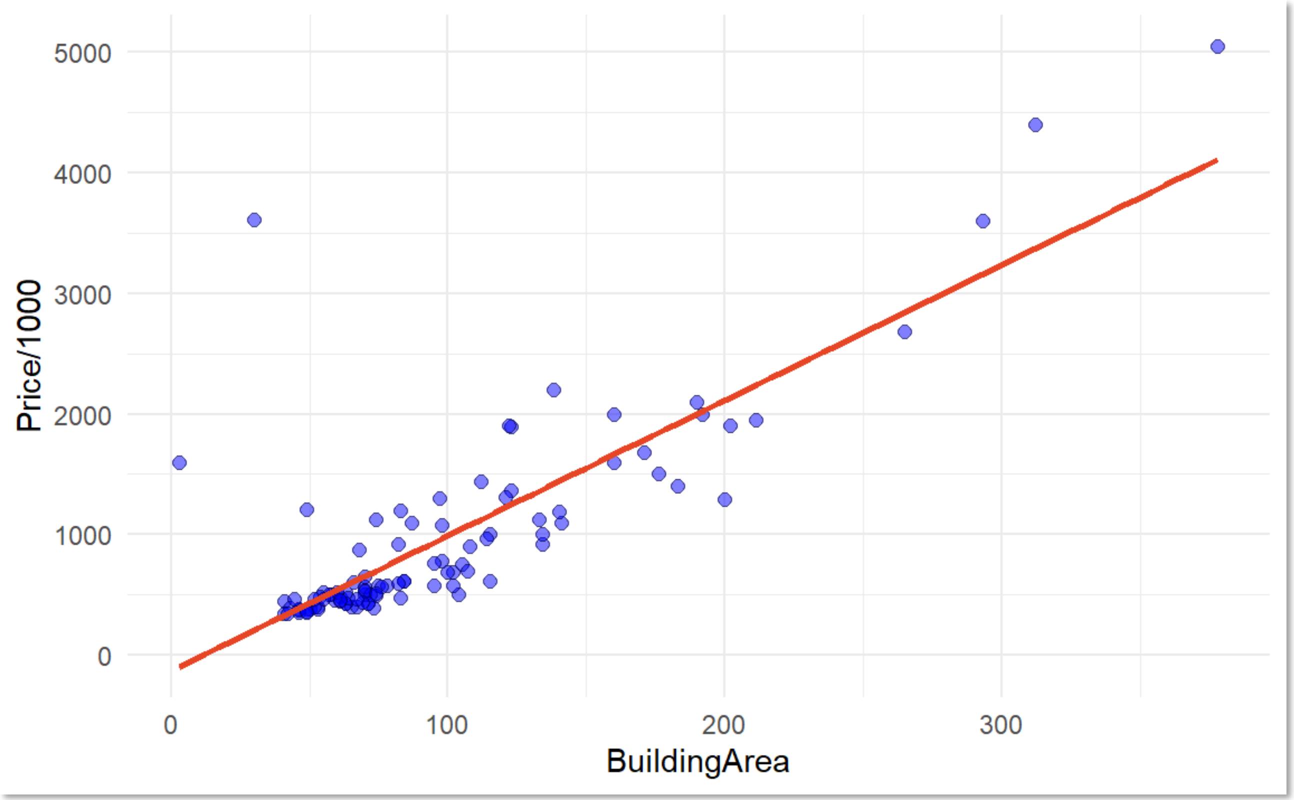
THE UNIVERSITY OF
SYDNEY

"extending Lin Reg"

1 feature → many features

Extending Simple Linear Regression

Linear regression fit



Recap: St Kilda house price linear model

```
lm.fit <- lm(Price ~ BuildingArea, data = st.kilda.data)
summary(lm.fit)
```

```
Call:
lm(formula = Price ~ BuildingArea, data = st.kilda.data)

Residuals:
    Min       1Q  Median       3Q      Max
-817415 -201614  -85181    19895 3403199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -129484.0    91775.9  -1.411    0.161
BuildingArea   11209.5      799.8  14.015   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 490300 on 99 degrees of freedom
Multiple R-squared:  0.6649,    Adjusted R-squared:  0.6615
F-statistic: 196.4 on 1 and 99 DF,  p-value: < 2.2e-16
```

$R^2 = 66\%$, remember again that we are looking for multiple R^2 in R, not adjusted.

Goodness of fit statistic

- Goodness of fit is measured by the **coefficient of determination** or R^2

when explaining R^2 always remember "proportion of variation in response Y"

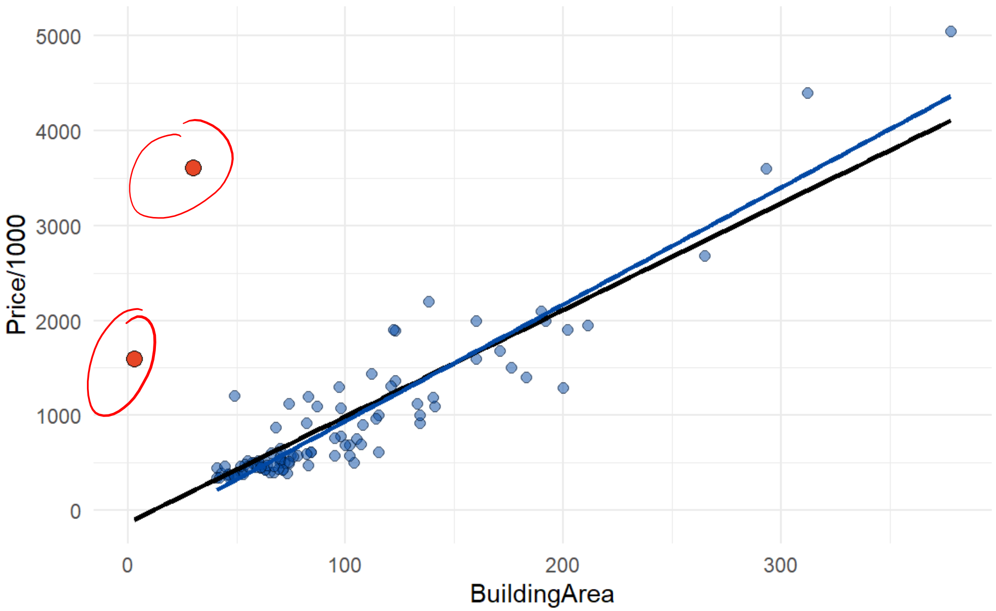
$$\boxed{R^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- It measures the proportion of variation in the response Y, explained by the linear regression on X very important sentence.
  - A value of 0 indicates **none** of the variance in Y can be explained linearly by X
  - A value of 1 indicates **all** of the variance in Y can be explained linearly by X

$0 \le R^2 \le 1$

Fit improvements

- Remove outliers:
- black line gives overall fit
- blue line fit only to blue data (without **red** points)



Linear fit after removing the outliers

```
lm.without.outliers <- lm(Price/1000 ~ BuildingArea, data = st.kilda.data, subset = BuildingArea >= 40)
summary(lm.without.outliers)
```

```
Call:
lm(formula = Price/1000 ~ BuildingArea, data = st.kilda.data,
    subset = BuildingArea >= 40)

Residuals:
    Min      1Q  Median      3Q     Max
-876.75 -137.30  -18.27  109.28  896.31

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -289.254     57.471  -5.033 2.22e-06 ***
BuildingArea   12.305      0.496  24.807  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 298.5 on 97 degrees of freedom
Multiple R-squared:  0.8638,    Adjusted R-squared:  0.8624
F-statistic: 615.4 on 1 and 97 DF,  p-value: < 2.2e-16
```

`R` formulae

additive

- Example formula

  $\boxed{\texttt{Response ~ Predictor1 + Predictor2 + Predictor3}}$
- Left hand side of ~ is the **response** variable (target to predict)
- Right hand side of ~ are the **predictor** variables (features)
- Relationship is assumed to be additive

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

- Interaction or multiplicative terms are denoted with `:` and `*` are beyond the scope for this course. e.g.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1 X_2 + \beta_3 X_2 + \cdots$$

Multiple linear regression

- Real life problems usually have more than one predictor.

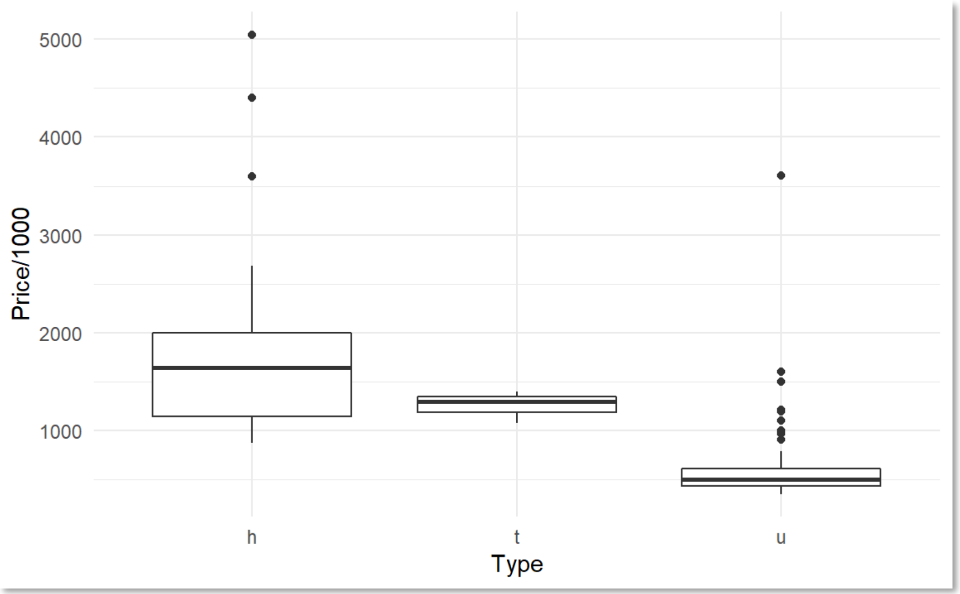
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p + \varepsilon$$

- Model fitting mathematics not discussed
  - Based on same principle (minimise residual sum of squares)
  - Uses multivariable calculus.

Extending model to multiple features

Perhaps 100 m^2 houses cost more than 100 m^2 units?

```
ggplot(st.kilda.data,
       aes(x = Type, y = Price/1000)) +
    geom_boxplot() +
    theme_minimal()
```



house townhouse unit

Multiple regression with `lm`

```r
multi.lm <- lm(Price/1000 ~ Type + BuildingArea, data = st.kilda.data)
summary(multi.lm)
```

```
Call:
lm(formula = Price/1000 ~ Type + BuildingArea, data = st.kilda.data)

Residuals:
   Min     1Q Median     3Q    Max 
-700.3 -173.1  -65.9   18.6 3389.6 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   342.865    186.764   1.836  0.06945 .  
Typet        -613.953    286.272  -2.145  0.03448 *  
Typeu        -408.417    139.915  -2.919  0.00436 ** 
BuildingArea    9.533      1.014   9.398 2.68e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 469.5 on 97 degrees of freedom
Multiple R-squared:  0.6989,	Adjusted R-squared:  0.6896 
F-statistic: 75.06 on 3 and 97 DF,  p-value: < 2.2e-16
```

$$\widehat{Price} = \begin{cases} 342 + 9.5\,Area, & \text{if house} \\ 342 - 613 + 9.5\,Area, & \text{if town house} \\ 342 - 408 + 9.5\,Area, & \text{if unit} \end{cases}$$

THE UNIVERSITY OF
SYDNEY

Interpreting Regression Models

Interpretation of Regression coefficients

- Simple case

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

  - β_1: the average change in Y for each unit increase in X_1.
  - β_0: the average of Y when $X_1 = 0$
- Multiple regression case

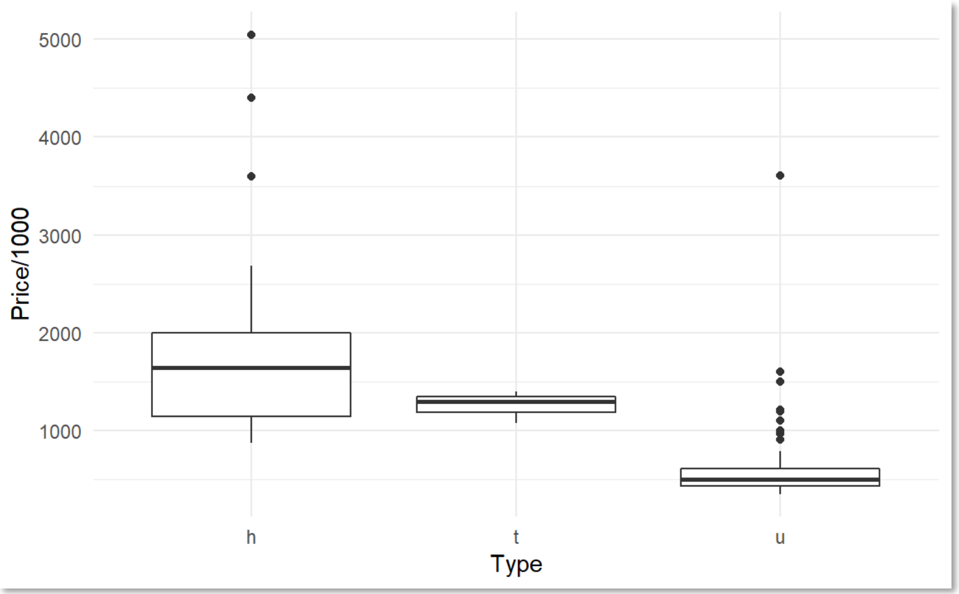
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p + \varepsilon$$

  - β_p: the average change in Y for each single unit increase in X_p ,holding all the other predictors fixed.

Extending model to multiple features

Perhaps 100 m^2 houses cost more than 100 m^2 units?

```
ggplot(st.kilda.data,
       aes(x = Type, y = Price/1000)) +
    geom_boxplot() +
    theme_minimal()
```



Multiple regression with `lm`

```r
multi.lm <- lm(Price/1000 ~ Type + BuildingArea, data = st.kilda.data)
summary(multi.lm)
```

```
Call:
lm(formula = Price/1000 ~ Type + BuildingArea, data = st.kilda.data)

Residuals:
   Min     1Q Median     3Q    Max
-700.3 -173.1  -65.9   18.6 3389.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   342.865    186.764   1.836  0.06945 .
Typet        -613.953    286.272  -2.145  0.03448 *
Typeu        -408.417    139.915  -2.919  0.00436 **
BuildingArea    9.533      1.014   9.398 2.68e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 469.5 on 97 degrees of freedom
Multiple R-squared:  0.6989,	Adjusted R-squared:  0.6896
F-statistic: 75.06 on 3 and 97 DF,  p-value: < 2.2e-16
```

Price is in the thousands

$$\widehat{Price} = \begin{cases} 342.8 + 9.5 \times Area, & \text{if house} \\ 342 - 613 + 9.5\, Area, & \text{if townhouse } \cancel{\text{unit}} \\ 342 - 408 + 9.5\, Area, & \text{if unit} \end{cases}$$

Model interpretation

```
summary(multi.lm)
...
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     342.865    186.764   1.836  0.06945 .
Typet          -613.953    286.272  -2.145  0.03448 *
Typeu          -408.417    139.915  -2.919  0.00436 **
BuildingArea      9.533      1.014   9.398 2.68e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
...
```

```
multi.pred.data <- data.frame(BuildingArea = rep(100, 3),
                              Type = c("u", "t", "h"))
predict(multi.lm, newdata = multi.pred.data)
```

```
        1         2         3
 887.7252  682.1894 1296.1427
```

newdata determines the predicted values.