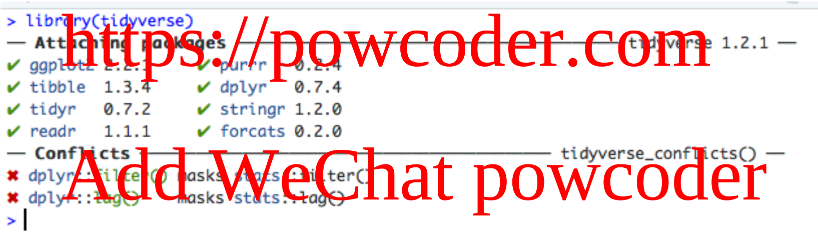
Lecture 5: Regression Part 1

Spatial Data Science II

Dr. Adams

Tidyverse

```r
library(tidyverse)
```



```
> library(tidyverse)
— Attaching packages ————————————————————————— tidyverse 1.2.1 —
✔ ggplot2 2.2.1     ✔ purrr   0.2.4
✔ tibble  1.3.4     ✔ dplyr   0.7.4
✔ tidyr   0.7.2     ✔ stringr 1.2.0
✔ readr   1.1.1     ✔ forcats 0.2.0
— Conflicts ———————————————————————————— tidyverse_conflicts() —
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
> 
```

Review

1. Examining data distributions
2. Data Visulization
3. Data Management
   - Tidy data

Big Bang

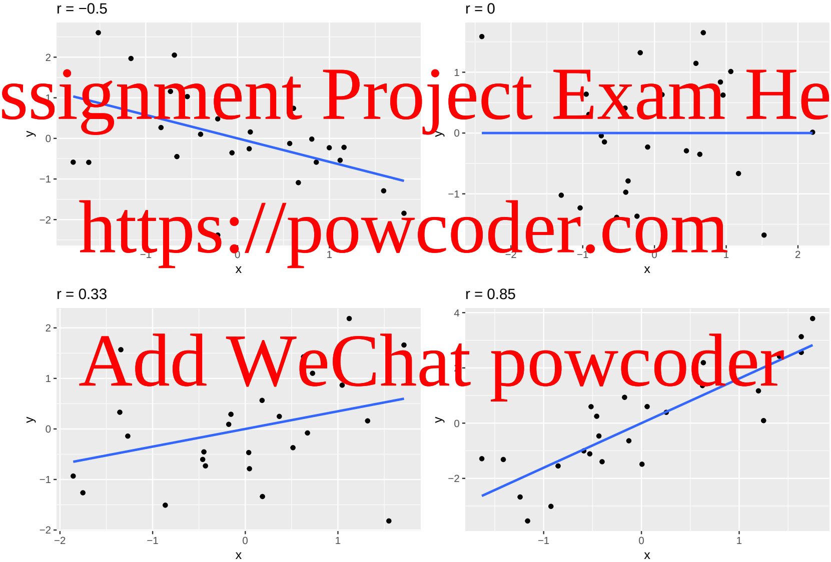
Correlation - Review

A measure of the dependence between two variables.

Pearson's product-moment coefficient

- Measure of the strength of a linear relationship between two variables.
- Coefficient is represented with r
- Ranges from -1 to $+1$
- $H_0 : r = 0$
- $H_1 : \neq 0$

Examples


r = −0.5

r = 0

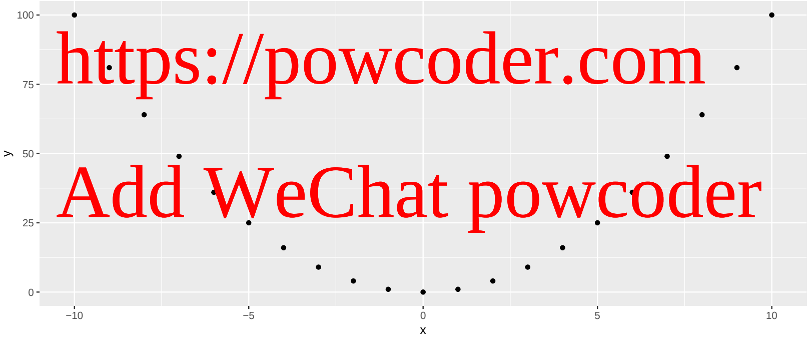
r = 0.33

r = 0.85

Pearson's *r* Assumptions

- Normally distributed
- Linear relationship

Linear assumption not met.



Spearman's rank correlation coefficient

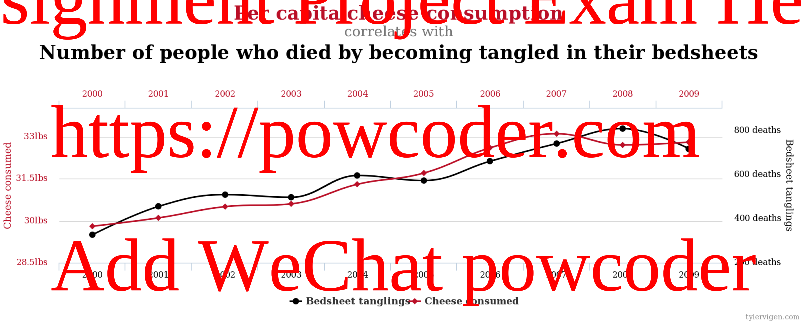
- Spearman's ρ or r_s
- Non-parametric (distribution-free) rank statistic
- Relationship does not need to be linear
- Accepts non-interval data

Spearman's ρ

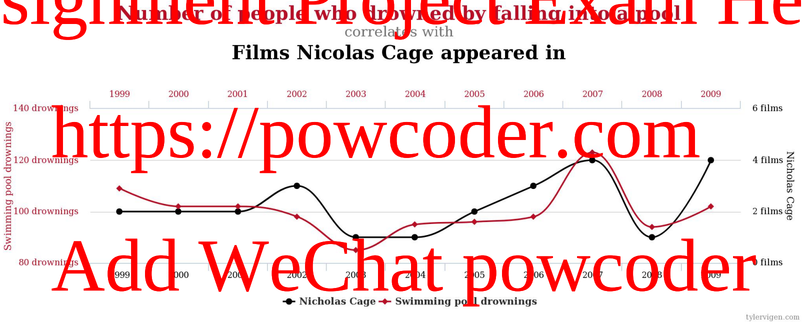
It is possible for r to be positive while ρ is negative.

"Make sure not to overinterpret Spearman's rank correlation coefficient as a significant measure of the strength of the associations between two variables" (Hauke and Kossowski 2011)

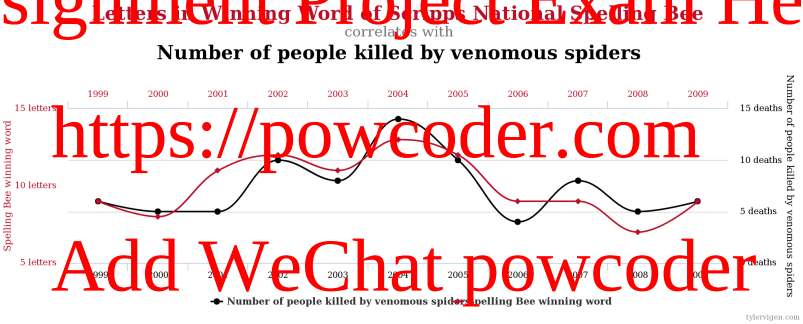
Spurious Correlations I



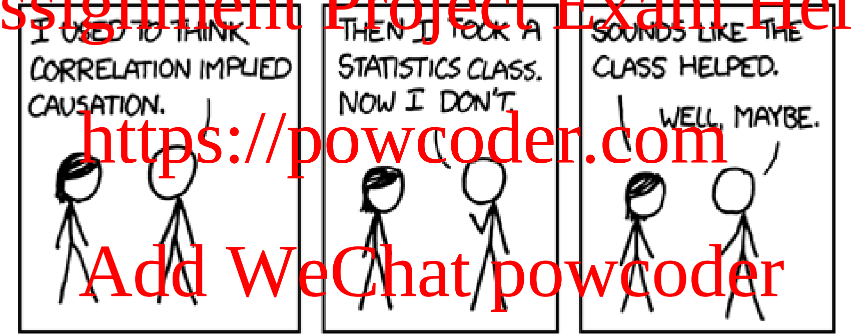
Spurious Correlations II



Spurious Correlations III



Correlation / Causation


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Moving On

How much does a house price increase when we increase its square footage?

- Take a minute and think about how you could answer this question?

Linear Regression Model

Model a continuous variable as a linear function of one or more independent variables.

- This allows us to understand if and how an attribute contributes to an outcome.

Model Use In Data Science

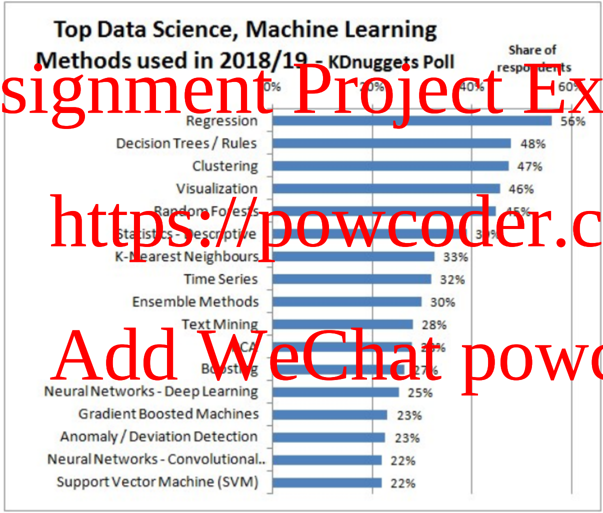


Fig. 1: Top Data Science, Machine Learning Methods Used, 2018/2019

Linear Regression

A statistical model that:

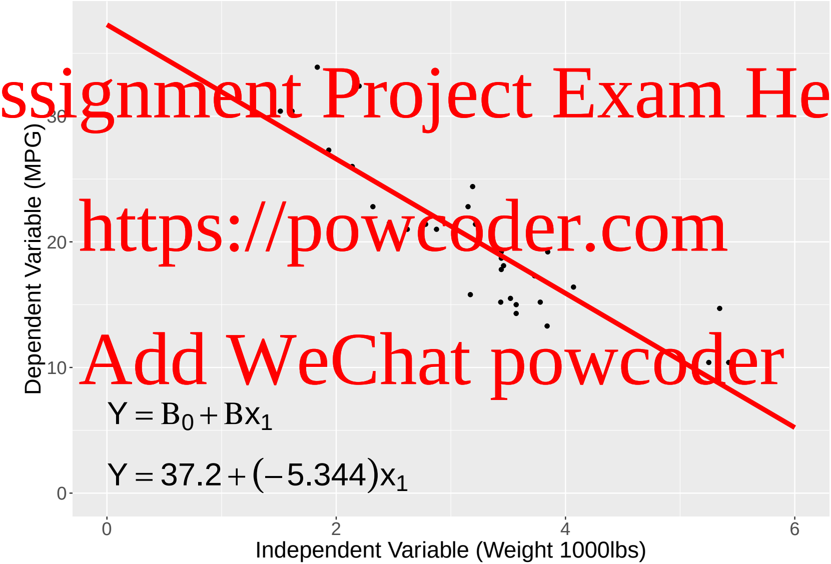
- Predicts a continuous variable: Y
- Using one or more independent variables: x_n
- Calculates a set of multipliers: β_n
  - Regression coefficients
- Includes an intercept: β_0

The linear regression formula is:

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 ... \beta_n x_n + \epsilon$

ϵ is our error or noise.

Linear Regression Visualized



$$Y = B_0 + Bx_1$$

$$Y = 37.2 + (-5.344)x_1$$

Housing Example - Explanatory

- House Price ($)
- Number of bedrooms
- Square Footage
- Number of bathrooms

House Sales Data

```
library(tidyverse)
house = read_delim(
  "http://www.rossmanchance.com/iscam2/data/housing.txt",
  delim = "\t") # Tab deliminated
```

```
## Parsed with column specification:
## cols(
##   sqft = col_integer(),
##   price = col_integer(),
##   City = col_character(),
##   bedrooms = col_integer(),
##   baths = col_double()
## )
```

Take a look at our data

```
## # A tibble: 83 x 5
##     sqft  price City            bedrooms baths
##    <int>  <int> <chr>              <int> <dbl>
##  1  3392 339000 Dublin                 3  2.10
##  2  4100 899900 pleasanton             4  3.00
##  3  3200 448641 Clayton                5  4.00
##  4  1436 239999 Moraga                 4  3.00
##  5  1944 377500 Antioch                3  2.00
##  6  1500 299900 Danville               3  2.50
##  7  1700 265000 El Dorado Hills        4  3.00
##  8  2507 449000 Shingle Springs        4  3.00
##  9  1580 439950 McKinleyville          3  2.00
## 10  1500 699888 Marina                 4  2.00
## # ... with 73 more rows
```

Housing Example - Explanatory

- House Price (price)
- Number of bedrooms (bedrooms)
- Square Footage (sqft)
- Number of bathrooms (baths)

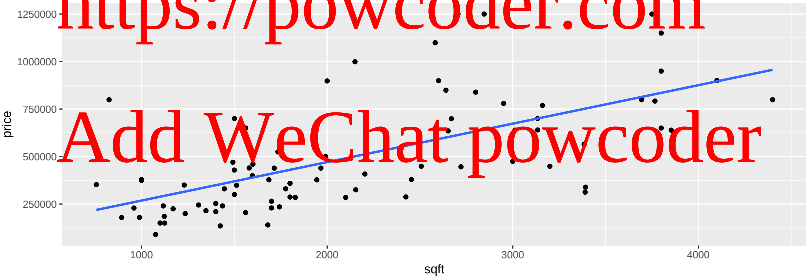
Data Validation

Before we fit our model we need to ensure the data fits the linear regression model's assumptions

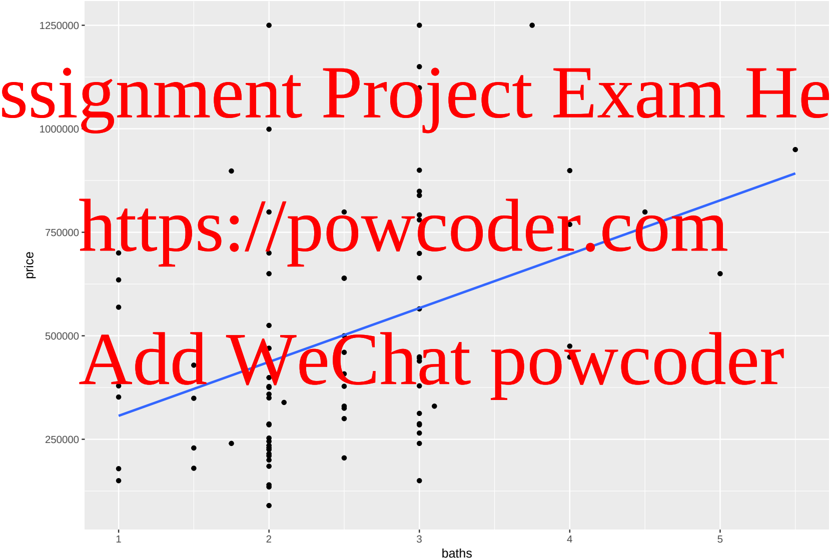
1. Linearity between independent variables (x_n) and dependent variable (Y)
2. No outliers in x_n
3. Normally Distributed x_n & Y

Linearity: Price and Square Footage

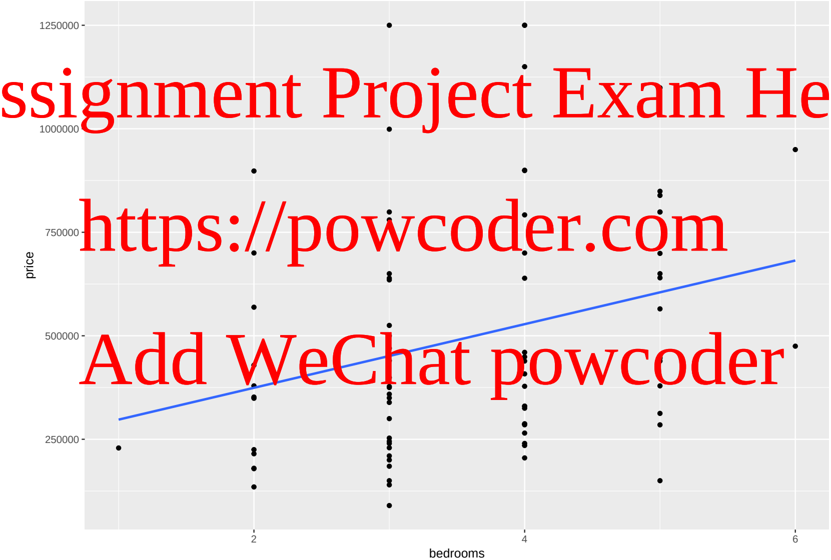
```r
ggplot(data = house, mapping = aes(y = price, x = sqft))+
  geom_point()+
  geom_smooth(method = "lm", se = F)
```



Linearity: Price and Baths



Linearity: Price and Bedrooms



Outliers - Box Plot

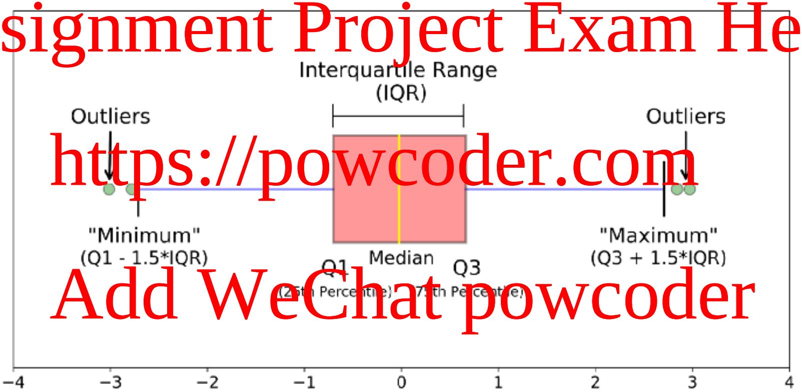
- Data values outside of 1.5 * interquartile-range may be considered outliers.
- The IQR is the distance from the 25th percentile to the 75th percentile.
- We often visualize this with the box and whisker plot

Box Plot

Box and whisker plots:

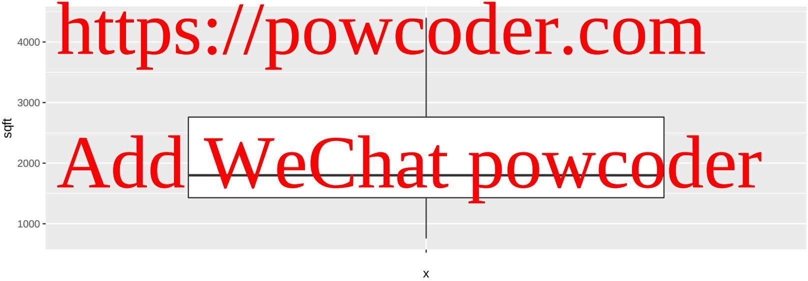
- bottom and top of the box are the first and third quartiles
- band inside the box is the second quartile (the median)
- geom_boxplot:
  - Whiskers are largest or smallest value within 1.5 * IQR
  - Points are outside of 1.5 * IQR

Box Plot Visualized



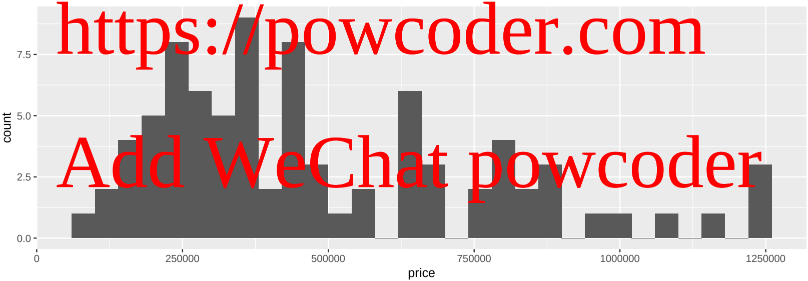
Check for outliers: Square Footage

```
ggplot(data = house)+
  geom_boxplot( mapping = aes(y = sqft, x = ""))
```



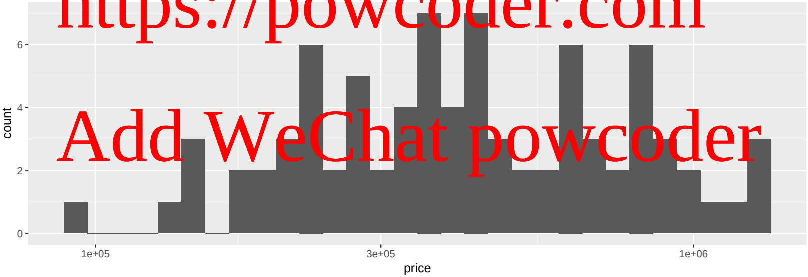
Distribution: Price

```
ggplot(data = house) +
  geom_histogram(mapping = aes(x = price))
```



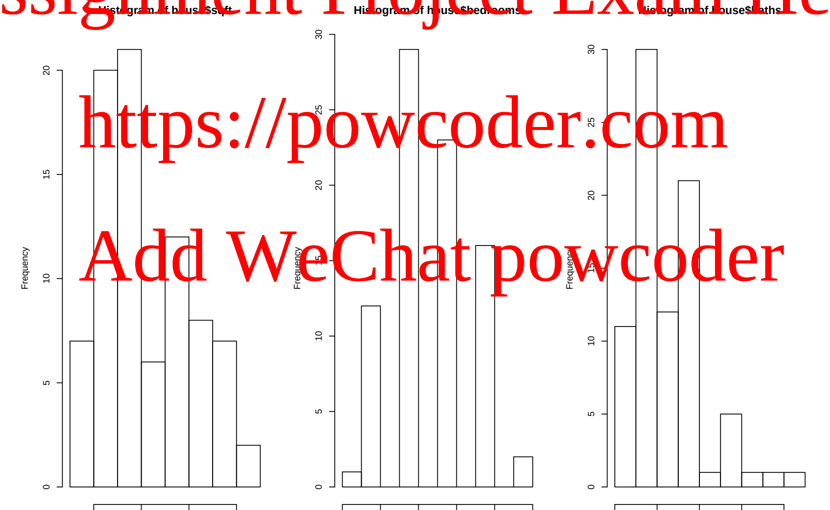
Distribution: log(Price)

```
ggplot(data = house)+
  geom_histogram(mapping = aes(x = price))+
  scale_x_log10()
```



Distribution: Independent Variables

```r
par(mfrow=c(1, 3))
hist(house$sqft); hist(house$bedrooms); hist(house$baths)
```



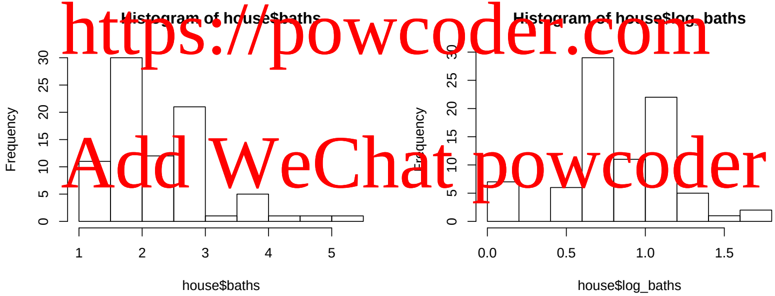
Transform Variables

```
house%>%
  mutate(price, log_price = log(price))%>%
  mutate(sqft, log_sqft = log(sqft))%>%
  mutate(baths, log_baths = log(baths)) -> house
house%>%
  select(log_baths)
```

```
## # A tibble: 83 x 1
##   log_baths
##       <dbl>
## 1     0.742
## 2     1.10
## 3     1.39
## 4     1.10
## 5     0.693
## 6     0.916
## 7     1.10
```

Transformed

```
par(mfrow=c(1, 2))
hist(house$baths); hist(house$log_baths)
```



Review: Graphical Analysis Pre-Check

- Scatter plot: Check for linear relationships between x_n and Y
- Box plot: Outlier check
- Histogram: Check variables for normal distributions

Explore Relationships: Correlation Strength

- It is good if a x_n is correlated with Y
- Problematic when multiple x_n are correlated
  - Multicollinearity
  - We will address later on in the lecture

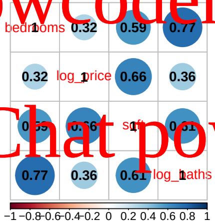
Correlation Matrix

```
cor_mat <- cor(house%>%
      select(bedrooms, log_price, sqft, log_baths))
cor_mat
```

```
##            bedrooms log_price      sqft log_baths
## bedrooms  1.0000000 0.3188020 0.5869470 0.7657905
## log_price 0.3188020 1.0000000 0.6638824 0.3554938
## sqft      0.5869470 0.6638824 1.0000000 0.6068155
## log_baths 0.7657905 0.3554938 0.6068155 1.0000000
```

Visualize Correlation: Plots

```
corrplot::corrplot(cor_mat, cl.pos = "b", tl.pos = "d", add
```



Tidyverse and modelling

- The tidyverse has yet to really address statistical modelling
- "You can see some of the pieces in the recipes and rsample packages but we do not yet have a cohesive system that solves a wide range of challenges. This work will largely replace the modelr package used in R4DS."- Tidyverse site
  - Broom is package that may be helpful.

Fit the linear regression model

- The function we use in R is stats::lm()
- The formula is:
  - dep ~ indep_1 + indep_2 + . . . + indep_n

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 ..\beta_n x_n + \epsilon$$

?lm: lm is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of co-variance (although aov may provide a more convenient interface for these).

R Formulas

Formulas use the tilde (by) and + (plus) characters:

$Y \sim var1 + var2 + var3 + \ldots varN$

Example:

```
lm(hwy~displ+year+cyl, data = mpg)
```

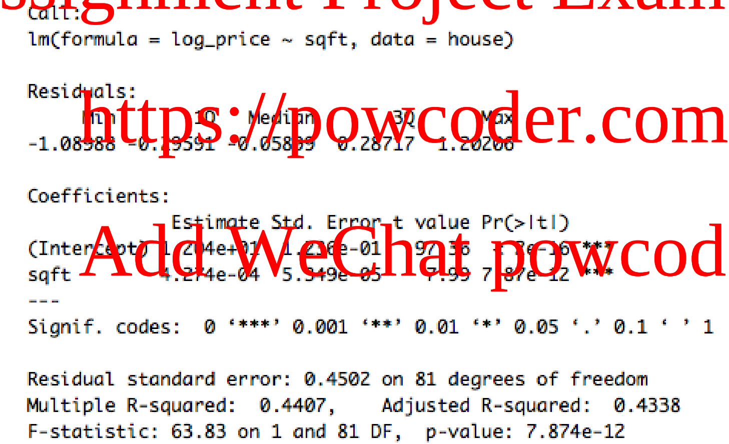
Linear Model

```
house_reg <- lm(log_price~sqft, data = house)
house_reg
```

```
##
## Call:
## lm(formula = log_price ~ sqft, data = house)
##
## Coefficients:
## (Intercept)         sqft
##   1.204e+01    4.274e-04
```

summary(lm())

```r
summary(house_reg)
```



```

Call:
lm(formula = log_price ~ sqft, data = house)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.08988 -0.29591 -0.05809  0.28717  1.20206 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.204e+01  1.236e-01   97.36  < 2e-16 ***
sqft        4.274e-04  5.349e-05    7.99 7.87e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4502 on 81 degrees of freedom
Multiple R-squared:  0.4407,	Adjusted R-squared:  0.4338 
F-statistic: 63.83 on 1 and 81 DF,  p-value: 7.874e-12
```

Accessing parts of the model

An object of class “lm” is a list containing at least the following components:

- coefficients: a named vector of coefficients
- residuals: the residuals, that is response minus fitted values.
- fitted.values: the fitted mean values.

```
house_reg$residuals[1:5]
```

Assessing the model

1. Are coefficients statistically significant?
   - Check with the coefficient *p-value*
   - Uses the t-value

2. Is the model statistically significant, overall p-value
   - Check with the coefficient *p-value*
   - Uses the F-test

Coefficients and t-value

The statistical signficance of each coefficicent is tested with the t-value

$t = \frac{coefficient}{std.error}$

- Should be greater 1.96 for p-value to be less than 0.05
- We reject the null hypothesis when $p < 0.05$
- When $p > 0.05$ we remove this variable from the model.

F-statistic

- The F-statistic assess the overall model.
- Null hypothesis:
  - An equal fit of the model with a model with zero predictors.
- Alternative hypothesis:
  - This model perform better than an intercept only model.
- If the p-value associated to the F-statistic is < 0.05 we reject H_0

Refresher on p-values if you need it

https://www.youtube.com/watch?v=128yz0OCG-I

Sidenote: Alpha, p-value, Type I and II errors

Type I Error:

- Incorrect rejection of a true null hypothesis
  - False positive

Type II Error:

- Incorrectly retaining a false alternative hypothesis
  - False negative

Alpha

As we decrease our chance of a Type I error, we increase our risk of Type II

R Squared & Adjusted R Squared

R^2 tells us is the proportion of variation in the dependent (response) variable that has been explained by this model.

The adjusted R^2 accounts for the effect that occurs when you add more independent variables that your R^2 increases.

- Increases only if a new term improves the model more than expected by chance.
- Decreases when a predictor improves the model by less than expected by chance.

Assumptions of the model

Checking your model:

1. The mean of the residuals is zero
2. Homoscedasticity of residuals or equal variance
3. Multicollinearity
4. The x_n variables and residuals are uncorrelated
5. The variability in X values is positive
6. The number of observations must be greater than number of X_n
7. Normality of residuals
8. No auto-correlation of residuals

Residuals Zero Mean

The mean of the residuals is zero. Residuals difference in model estimates and actual values.

```
mean(house_reg$residuals)
```

```
## [1] 2.70306e-17
```

Homoscedasticity: Equal variance across values

Homoscedasticity:

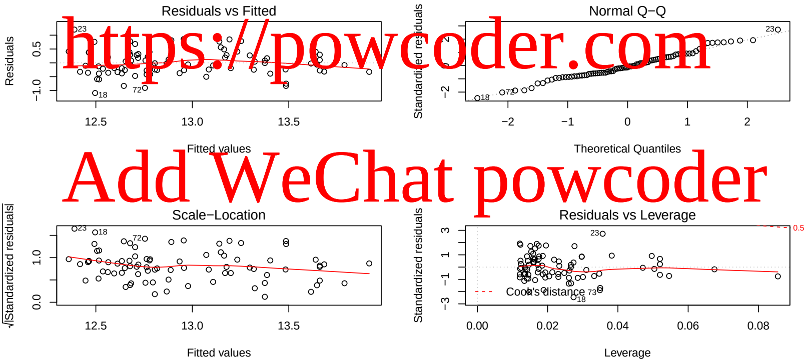
- Requires variance of residuals to be the same across the fitted values.

Heteroscedasticity:

- When the size of the error term differs across values of an independent variable.
  - Violation of homoscedasticity
- Linear Regression (Ordinary Least Squares), seeks to minimize residuals
- OLS equally weights all observations
  - Cases with larger errors have more effect on the model estimation.

Checking with plot(lm(y∼x, data = data))

```
par(mfrow=c(2,2)) # set 2 rows and 2 column plot layout
plot(house_reg)
```



Dealing with Heteroscedasticity

```
# Breusch-Pagan test
lmtest::bptest(house_reg)
car::ncvTest(house_reg)
```

1. Try different predictors
2. Variable transformation
   - Box-Cox
3. Select a different regression model (last case)

```
library(caret)
bc <- BoxCoxTrans('values')
predict(bc, 'values')
```

No autocorrelation of residuals

- Temporal Autocorrelation
  - The value at one point is not dependent on the previous value
- Spatial Autocorrelation
  - Values at one location are not dependent on near values
    - Moran's I

The x_n variables and residuals are uncorrelated

```r
cor.test(house$sqft, house_reg$residuals)
```

```
##
##  Pearson's product-moment correlation
##
## data:  house$sqft and house_reg$residuals
## t = -4.7743e-16, df = 81, p-value = 1
## alternative hypothesis: true correlation is not equal to
## 95 percent confidence interval:
##  -0.2156893  0.2156893
## sample estimates:
##           cor
## -5.304794e-17
```

The variability in x_n values is positive

```
var(house$sqft)
```

```
## [1] 863896.6
```

This is much greater than 0.

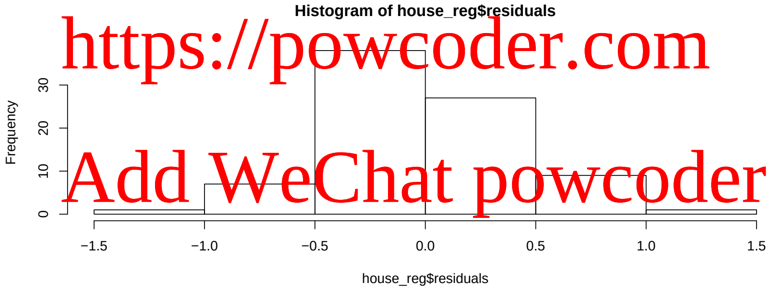
We cannot use a x_n variable with a single value.

The number of observations must be greater than number of Xs

Unlikely to be an issue except with extreme cases.

Normality of residuals

```
hist(house_reg$residuals)
```



No perfect multicollinearity

Occurs when we have 2 or more predictor variables.

- Assessed using VIF
  - Variance inflation factors

- Rule of thumb VIF > 4
  - You should revise your variable selection

```
cars::vif()
```

Motor Trend Car Road Tests

```r
car::mtcars
```

- mpg: Miles/(US) gallon
- cyl: Number of cylinders
- disp: Displacement (cu.in.)
- hp: Gross horsepower
- drat: Rear axle ratio
- wt: Weight (1000 lbs)
- qsec: 1/4 mile time
- vs: V/S
- am: Transmission (0 = automatic, 1 = manual)
- gear: Number of forward gears
- carb: Number of carburetors

Linear Model for mpg

```
library(car)
# We will include all variables in the model
mpg_lm <- lm(mpg ~ ., data=mtcars)
vif(mpg_lm)
```

```
##       cyl      disp        hp      drat        wt         q
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527
##        am      gear      carb
##  4.648487  5.357452  7.908747
```

Remove variables in mpg Model

```r
mpg_lm_2 <- lm(mpg ~ cyl + gear + am, data=mtcars)
vif(mpg_lm_2)
```

```
##      cyl     gear       am
## 1.407382 2.768828 2.884543
```

Issues caused by multicollinearity

- Increase the variance of the coefficient estimates
- Estimates may be very sensitive to minor changes in the model
- Statistical power is reduced
- Coefficient sign switching

"A caution must be taken when more than two predictors in the model have even weak pairwise correlation coefficients (r=0.25) as they can result in a significant multicollinearity effect." (P. Vatcheva and Lee 2016 ,pg. 13)

References

Hauke, Jan, and Tomasz Kossowski. 2011. "Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data." *Quaestiones Geographicae* 30 (2): 87–93. doi:10.2478/v10117-011-0021-1.

P. Vatcheva, Kristina, and MinJae Lee. 2016. "Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies." *Epidemiology: Open Access* 06 (02): 1–20. doi:10.4172/2161-1165.1000227.