LECTURE 4 - CORRELATION & REGRESSION – PART B

Multiple Regression

LECTURE 4B-1 – CORRELATION AND PARTIAL CORRELATION & MULTIPLE REGRESSION

Multiple Regression

- Multiple regression involves multiple predictors (independent variables) predicting a single dependent variable.
- Population Model:
  - $Y = \boldsymbol{\alpha} + \boldsymbol{\beta_1} X_1 + \boldsymbol{\beta_2} X_2 + \boldsymbol{\beta_3} X_3 + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\sigma_\varepsilon}^2)$.
- (Sample) Regression Model or Prediction Model
  - $\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$ with residual (also called error or noise) terms $\boldsymbol{e} = (y - \hat{y})$
- Standardized (sample) Model
  - $\hat{z}_y = \hat{a} + \hat{b}_1 z_{X1} + \hat{b}_2 z_{X2} + \hat{b}_3 z_{X2}$

Dataset – MR100.csv; **MultipReg.R**

- For Multiple Regression, we will use a larger data set with 100 observations
- The data set is available as a downloadable Data Set MR100.csv from Canvas.
- Our goal is to develop a multiple regression model that predicts Assets using Age, Home Value and Mortgage.

Obs	Gender	Marital_Status	Age	Home_Value	Mortgage_Balance	Assets
1	0	1	71	498.68	156.33	142.49
2	1	1	60	16.21	47.43	117.29
3	1	1	82	427.71	127.7	134.15
4	1	1	87	113.15	191.01	139.08
5	0	0	21	0.07	15.58	100.01
6	0	0	31	0.17	13.85	85.72
7	0	1	86	11.58	36.61	106.2
8	0	1	37	0.08	2.49	64.02
9	0	1	56	25.2	106.63	108.16
10	0	1	62	4.92	34.49	97.62
11	0	1	95	7.01	17.14	82.68
12	0	0	34	1.04	12.89	87.33
13	0	0	64	2.79	71.69	126.67
14	0	0	64	3.57	43.18	87.73
15	1	1	50	1.07	15.23	72.81
16	0	1	40	0.04	0.76	43.93
17	0	1	64	21.59	8.21	83.81
18	0	1	64	44.84	94.18	143.63
19	0	1	77	5.27	128.18	135.06
20	1	1	71	0.19	11.93	95.27
21	1	1	56	0.09	0.8	58.77
22	0	1	19	1.22	5.78	69.31
23	0	1	57	24.84	134.13	110.59
24	0	0	60	0.25	3.95	82.75
25	0	1	36	0.94	43.36	98.88
26	0	1	47	0.87	83.12	115.88
27	0	0	39	2091.4	138.78	129.77
28	0	1	63	40.69	71.89	112.83
29	0	1	52	0.2	26.12	109.37
30	1	1	85	27.42	89.91	131.68
31	0	1	26	5.15	59.47	115.4
32	0	1	64	209.91	178.17	150.13
33	0	0	70	61.98	79.85	127.67

```
> # Clear the Environment
> rm(list=ls())
>
> # Read csv file as a DataFrame
> #
> setwd("C:\\Users\\sarathy\\Documents\\2019-Teaching\\Fall2019\\Fall2019-MSIS5503\\MSIS-5503-Data")
> df <- read.table('MR100.csv',
+                  header = TRUE, sep = ',')
>
> #Assign variable names to DataFrame Column objects
> id <- df$Obs
> gender <-df$Gender
> marital <- df$Marital_Status
> age<-df$Age
> home <- df$Home_Value
> mortgage<- df$Mortgage_Balance
> assets <-df$Assets
> #
> data <- cbind(age,home, mortgage, assets)
```

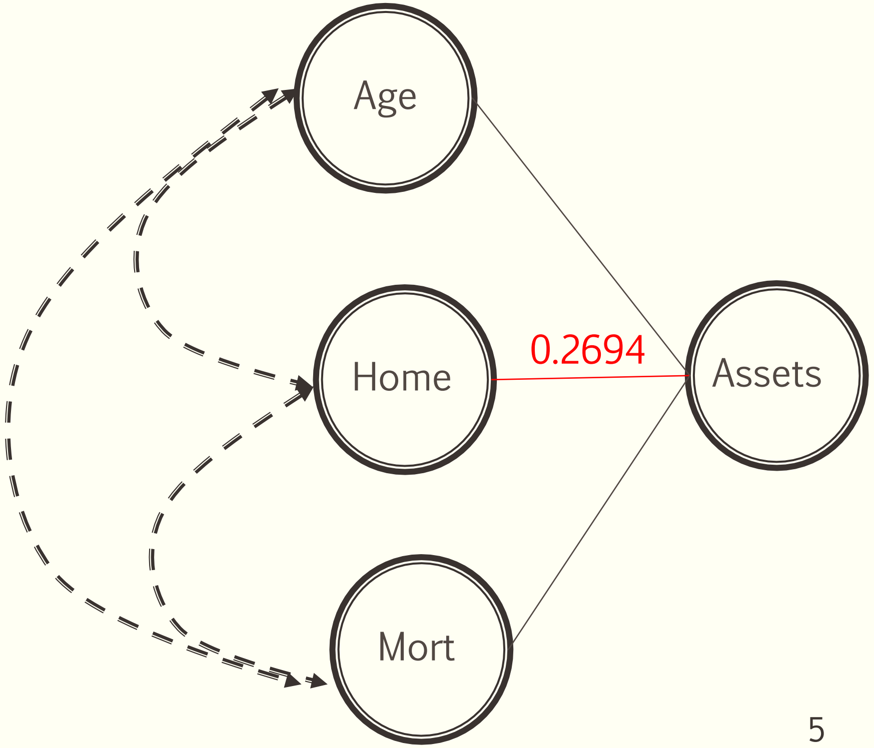
Correlation, Partial Correlation and Regression

- You can get the pairwise correlation matrix using:

```
> data <- cbind(age,home, mortgage, assets)
> corr <- cor(data)
> print(signif(corr),4)
```

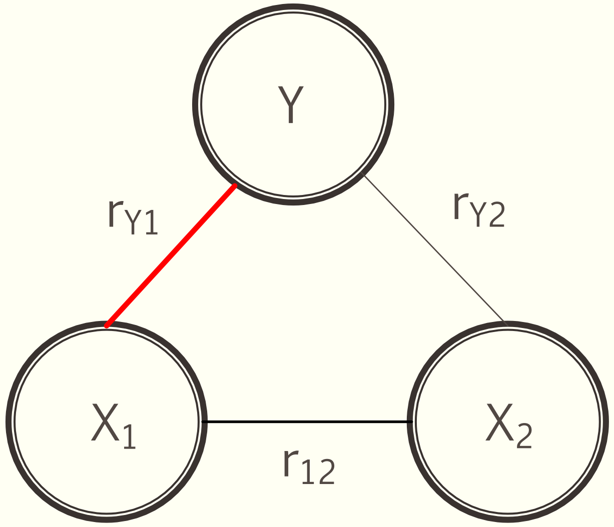
```
> corr <- cor(data)
> print(signif(corr),4)
              age    home mortgage assets
age       1.00000 0.04033   0.3990 0.4846
home      0.04033 1.00000   0.4139 0.2694
mortgage  0.39899 0.41392   1.0000 0.7495
assets    0.48456 0.26938   0.7495 1.0000
```

- The correlation matrix seems to suggest that the correlation between home value and assets is 0.2694, based on pairwise correlations.
- But, the pairwise correlation **does not** capture the *effect (correlations) of other variables on both home value or assets.*
- We want to look at the "true" correlation between home value and assets
- That is, we want to "remove" (or "control for") the effect of other variables on the correlation relationship between home value and assets.
- One way to do this is through *partial correlations.*



Partial Correlation

- Consider two variables X_1 and X_2 that are used to predict a dependent variable Y.
- Let
  - r_{Y1} be the correlation between Y and X_1,
  - r_{Y2} be the correlation between Y and X_2, and
  - r_{12} be the correlation between X_1 and X_2.
- The formula for:
  - The partial correlation coefficient between X_1 and Y, controlling for the correlation between (X_1 and X_2) and (X_2 and Y), are shown to the right.
  - You can see that the numerator "removes (controls or partials out) the correlation between Y and X_2 (r_{Y2}) and between X_1 and X_2 (r_{12}) when looking at the partial correlation between Y and X_1 (r_{Y1})
- The partial correlation formula, when we control for more than one variable is more complicated, but can be obtained using **R** using the *pcor()* function.



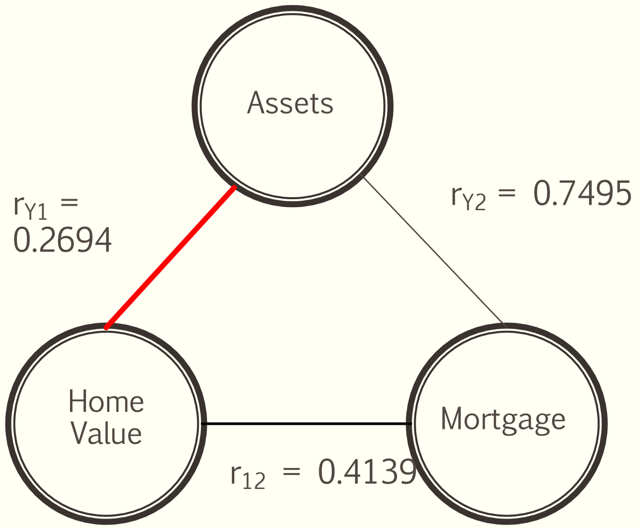
$$Partial = \frac{r_{y1} - (r_{y2})(r_{12})}{\sqrt{1 - r_{y2}^2}\sqrt{1 - (r_{12})^2}}$$

Partial Correlations

```
#
# Calculate correlations using cor() and partial correlations using pcor()
# for Home Value, Mortgage and Assets
#
# install.packages(ppcor)
library(ppcor)
pdata <- cbind(home, mortgage, assets)
corr <- cor(pdata)
print(signif(corr, 4))
pcorr <- pcor(pdata)
print(pcorr)
```

- In **R**, make sure you install the “ppcor” package, using install.packages(“ppcor”). You do this only once.
- Then attach the library using *library(ppcor)*
- The partial correlation function is *pcor()*
- You only want the *$estimate* portion of the output
-

```
> corr <- cor(pdata)
> print(signif(corr, 4))
           home mortgage assets
home     1.0000   0.4139 0.2694
mortgage 0.4139   1.0000 0.7495
assets   0.2694   0.7495 1.0000
> pcorr <- pcor(pdata)
> print(pcorr)
$estimate
                home  mortgage      assets
home      1.00000000 0.3325614 -0.06779048
mortgage  0.33256141 1.0000000  0.72775873
assets   -0.06779048 0.7277587  1.00000000
```



- The *partial correlation* between Assets and Home Value, controlling for the correlation of Mortgage on both of them is *-0.068* compared to the pairwise (uncontrolled) correlation of 0.2694 between Assets and Home Value!

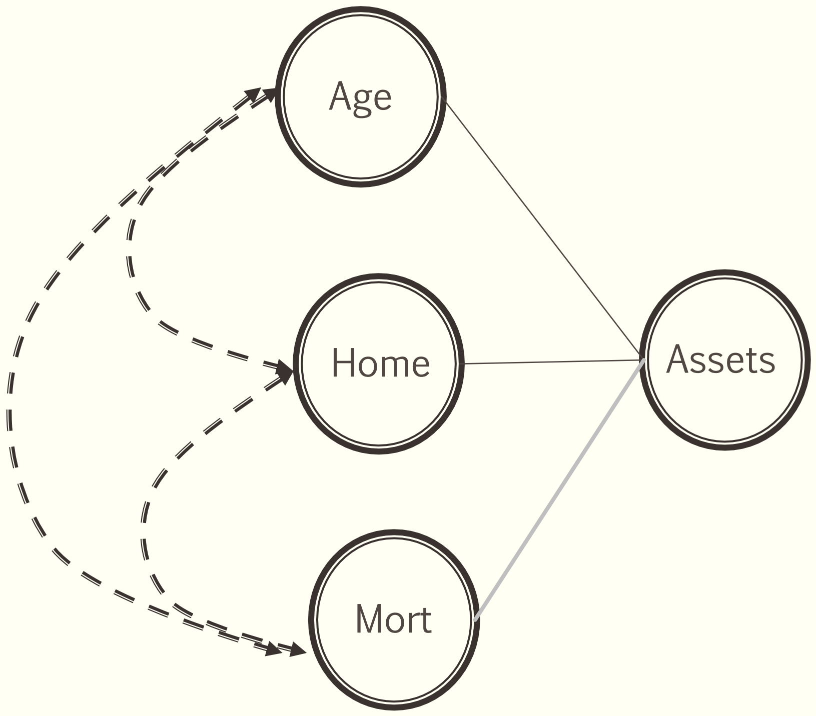
$$Partial = \frac{r_{y1} - (r_{y2})(r_{12})}{\sqrt{1 - r_{y2}^2}\sqrt{1 - (r_{12})^2}}$$

```
> pc = (0.2694 - (0.4139*0.7495))/((sqrt(1-0.4139^2)*sqrt(1-0.7495^2)))
> pc
[1] -0.06773236
```

- It tells us that the correlation matrix does not give a true picture of the linear relationship between Assets and Home Value.

Multiple Regression

- Multiple regression involves multiple predictors (independent variables) predicting a single dependent variable.
- Population Model:
  - $Y = \boldsymbol{\alpha} + \boldsymbol{\beta_1} X_1 + \boldsymbol{\beta_2} X_2 + \boldsymbol{\beta_3} X_3 + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\sigma_\varepsilon}^2)$.
- (Sample) Regression Model or Prediction Model
  - $\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$
  - The residual (also called error or noise) terms $\boldsymbol{e} = (y - \hat{y})$
- Standardized (sample) Model
  - $\hat{z}_y = \hat{a} + \hat{b}_1 z_{X1} + \hat{b}_2 z_{X2} + \hat{b}_3 z_{X2}$



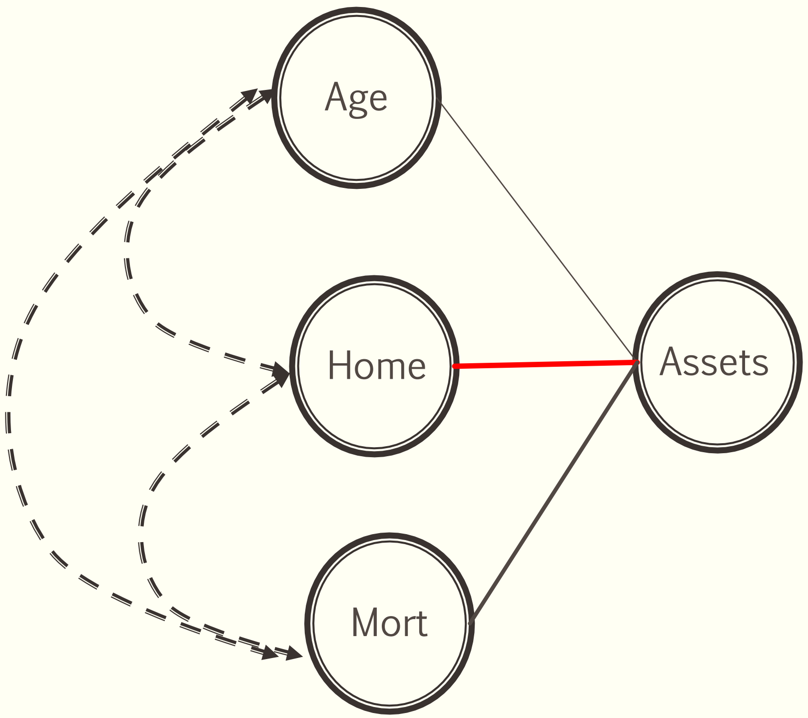
Partial Correlation and Multiple Regression

- We saw earlier that Partial correlation is the correlation between two variables after *controlling both for the effect of common variables*. In our case, we could compute the partial correlation between Assets and Home Value, controlling for the correlation of Mortgage on both of them

- However, the problem with partial correlation is that when we have 4 variables or more, controlling for the effect of all other variables on any pair becomes tedious and complicated.

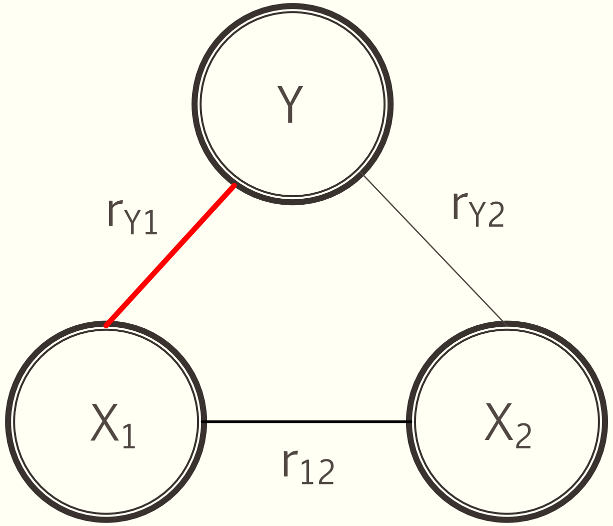
- We would have to do this for every pair of variables, controlling for the effect of all other variables on both of them.

- Fortunately, multiple regression can do the same task (controlling each predictor for the effect of other predictors on the independent variable) much more easily



Correlation, Partial Correlation and Regression

- Consider two variables X_1 and X_2 that are used to predict a dependent variable Y.
- Let
  - r_{Y1} be the correlation between Y and X_1,
  - r_{Y2} be the correlation between Y and X_2, and
  - r_{12} be the correlation between X_1 and X_2.
- The formulas for:
  - The slope coefficient β_1 when X_1 and X_2 are used in a regression model to predict Y, and
  - The partial correlation coefficient between X_1 and Y, controlling for the correlation between X_1 and X_2 are shown to the right.
- You can see that the regression $\boldsymbol{\beta_1}$ also "removes" the effect of the correlation between X_2 and X_1 when calculating the relationship between X_1 and Y through the term $r_{Y1} - (r_{Y2})(r_{12})$
- Thus, when we perform regression and look at the slope ($\boldsymbol{\beta_1}$) of X_1,
  - we are looking at the linear relationship between X_1 and the Y, *controlling for the correlation of all other predictors with X_1 and Y.*
  - That is, ($\boldsymbol{\beta_1}$) is the amount of increase in Y for a unit change in X_1, *controlling for the correlation of all other predictors with X_1 and Y.*
  - Another way to say this is, ($\boldsymbol{\beta_1}$) is the amount of increase in Y for a unit change in X_1, *holding all other predictors constant or fixed.*



$$\beta_1 = \frac{r_{y1} - (r_{y2})(r_{12})}{1 - (r_{12})^2}$$

$$Partial = \frac{r_{y1} - (r_{y2})(r_{12})}{\sqrt{1 - r_{y2}^2}\sqrt{1 - (r_{12})^2}}$$

Multiple Regression – Controlling for the Effect of Other X variables

- In **R**:
  - m_reg1 <- lm(assets ~ age+home+mortgage)
  - summary(m_reg1)
- The model:
  - $\widehat{\text{assets}} = 70.00 + 0.270\ \text{age}\ \mathbf{-0.001}\ \text{homevalue} + 0.327$ mortgage balance
- Like partial correlation, even though the correlation between Home Value and Assets was 0.269, *Once w* *control for the effect of age and mortgagebalance on homevalue*, the relationship between homevalue and assets practically disappears! (-0.001).
- This shows that regression coefficients are very similar to partial correlations, in controlling for the effect of the other predictor (independent variables) variables.

```
> #
> # Multiple Regression - Predict Assets using Age, Home Value and Mortgage Balance
> #
> m_reg1 <- lm(assets ~ age+home+mortgage)
> summary(m_reg1)

Call:
lm(formula = assets ~ age + home + mortgage)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.83 -10.42    2.36   10.02   33.07 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 70.004077   4.547547  15.394  < 2e-16 ***
age          0.269996   0.087776   3.076  0.00273 ** 
home        -0.001751   0.007616  -0.230  0.81861    
mortgage     0.327100   0.038056   8.595 1.54e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.44 on 96 degrees of freedom
Multiple R-squared:  0.6029,	Adjusted R-squared:  0.5905 
F-statistic: 48.58 on 3 and 96 DF,  p-value: < 2.2e-16
```

```
> corr <- cor(data)
> print(signif(corr),4)
              age    home mortgage assets
age       1.00000 0.04033   0.3990 0.4846
home      0.04033 1.00000   0.4139 0.2694
mortgage  0.39899 0.41392   1.0000 0.7495
assets    0.48456 0.26938   0.7495 1.0000
```

Multiple Regression – Significance Tests for Predictors and Interpretation of Betas

- **Age:**
  - Since the p-value for the t-test of $H_0{:}\beta_{age} = 0$ is 0.0027, we **reject** the null hypothesis at $\alpha = 0.05$.
  - We conclude that Age is a significant predictor of assets when Home Value and Mortgage Balance are included in the model.
    - Predicted Assets increase by 0.276 (1000's dollars) for each unit (year) increase in age, *controlling for* (or *holding constant*) the effect of Home Value and Mortgage Balance in the model.

- **Home Value:**
  - Since the p-value for the t-test of $H_0{:}\beta_{home} = 0$ = 0.8186, we **fail to reject** the null hypothesis at $\alpha = 0.05$
  - We conclude that Home Value is NOT a significant predictor of Assets when Age and Mortgage Balance are included in the model.

- **Mortgage Balance:**
  - Since the p-value for the t-test of $H_0{:}\beta_{mortgage} = 0$ is 0.000, we **reject** the null hypothesis $H_0{:}\beta_{mortgage} = 0$ at $\alpha = 0.05$,
  - We conclude that Mortgage Balance is a significant predictor of Assets when Home Value and Mortgage Balance are included in the model.
    - Predicted Assets increase by 0.327 (1000's dollars) for each unit (year) increase in Mortgage Balance *controlling for* (or *holding constant*) the effect of Home Value and Age in the model.

```
> #
> # Multiple Regression - Predict Assets using Age, Home Value and Mortgage Balance
> #
> m_reg1 <- lm(assets ~ age+home+mortgage)
> summary(m_reg1)

Call:
lm(formula = assets ~ age + home + mortgage)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.83  -10.42    2.36   10.02   33.07 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 70.004077   4.547547  15.394  < 2e-16 ***
age          0.269996   0.087776   3.076  0.00273 ** 
home        -0.001751   0.007616  -0.230  0.81861    
mortgage     0.327100   0.038056   8.595 1.54e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.44 on 96 degrees of freedom
Multiple R-squared:  0.6029,	Adjusted R-squared:  0.5905 
F-statistic: 48.58 on 3 and 96 DF,  p-value: < 2.2e-16
```

$$\widehat{assets} = 70.00 + 0.270\ age\ \mathbf{-0.001}\ homevalue + 0.327\ mortgage\ \ balance$$

Predictor Significance depends on other Predictors in the Model

- The significance (or lack of) significance of a predictor depends on what other predictors are included in the model.
- To see this, we run a simple regression predicting Assets with only Home Value in the model.
- You can clearly see that since the p-value for Home Value is 0.00872, without Mortgage Balance and Age, Home Value is a significant predictor at $\alpha = 0.05$.
- The R^2 is exactly the square of the correlation of Home Value with Assets.
- This is why we call the correlations in the correlation matrix as uncontrolled or *zero-order* correlation. Zero-order means that none of the effects of other variables have been partialled out, or controlled for, from the correlation between Home Value and Assets.
- This is also the reason why, in concluding about the significance of a predictor or interpreting it, *all the other predictors in the model have to be explicitly acknowledged.*

```
> # Simple Regression - Predict Assets using only Home Value
> #
> m_reg2 <- lm(assets ~ home)
> summary(m_reg2)

Call:
lm(formula = assets ~ home)

Residuals:
    Min      1Q  Median      3Q     Max 
-55.428 -14.440   3.473  15.828  44.747 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 99.35657    2.40173  41.369  < 2e-16 ***
home         0.02871    0.01037   2.769  0.00672 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.35 on 98 degrees of freedom
Multiple R-squared:  0.07257,	Adjusted R-squared:  0.0631 
F-statistic: 7.668 on 1 and 98 DF,  p-value: 0.006724
```

```
> corr <- cor(data)
> print(signif(corr),4)
             age    home mortgage assets
age      1.00000 0.04033   0.3990 0.4846
home     0.04033 1.00000   0.4139 0.2694
mortgage 0.39899 0.41392   1.0000 0.7495
assets   0.48456 0.26938   0.7495 1.0000
```

Comparing the Two Population Models

- Simple Regression Model:
  - Assets = $\boldsymbol{\alpha}_1$ + $\boldsymbol{\beta}_{\mathbf{Home}}$Home + $\boldsymbol{\epsilon}_1$ with $\boldsymbol{\epsilon}_1 \sim N(\mathbf{0}, \boldsymbol{\sigma}_{\varepsilon 1}^2)$.
  - Note: $\boldsymbol{\beta}_{\mathbf{Home}}$ was found to be significantly different from 0 at $\alpha = 0.05$.
- Multiple Regression Model:
  - Assets = $\boldsymbol{\alpha}_2$ + $\boldsymbol{\beta}_{\mathbf{Home}}$Home + $\boldsymbol{\beta}_{\mathbf{Age}}$Age + $\boldsymbol{\beta}_{\mathbf{Mortgage}}$Mortgage + $\boldsymbol{\epsilon}_2$ with $\boldsymbol{\epsilon}_2 \sim N(\mathbf{0}, \boldsymbol{\sigma}_{\varepsilon 2}^2)$.
  - Note: $\boldsymbol{\beta}_{\mathbf{Home}}$ was found to be NOT significantly different from 0 at $\alpha = 0.05$.
- In the first model, Age and Mortgage were "left in the error term" $\boldsymbol{\epsilon}_1$
- They are "brought into the model" in the second model, and in their presence Home is no longer a significant predictor of Assets.

Multiple Regression –Overall Model Fit F-Test

- The Null hypothesis for the Model Fit F-test in multiple regression is:
    - H_0: $\boldsymbol{\beta}_{\mathbf{age}} = \boldsymbol{\beta}_{\mathbf{home}} = \boldsymbol{\beta}_{\mathbf{mort}} = 0$
  - vs the alternate hypothesis:
    - H_a: At least one of $\boldsymbol{\beta}_{\mathbf{age}}, \boldsymbol{\beta}_{\mathbf{home}}, \boldsymbol{\beta}_{\mathbf{mort}} \neq 0$
  - is **rejected** at $\alpha = 0.05$ with a p-value of 0.000
- The F-statistic calculated as:
  - $MS_{Model} / MS_{Error} = 48.58$ with numerator degrees of freedom = 3 and denominator degrees of freedom (n-k-1 = 96)
  - k=3 is the number of predictors and n = 100 the sample size.
- The p-value for the F-statistic 0.000
- <u>**Conclusion**</u>: At least one of $\boldsymbol{\beta}_{\mathbf{age}}, \boldsymbol{\beta}_{\mathbf{home}}, \boldsymbol{\beta}_{\mathbf{mort}} \neq 0$ i.e., *at least one of the three predictors* Age, Home Value and Mortgage Balance is useful in predicting Assets.

```
> #
> # Multiple Regression - Predict Assets using Age, Home Value and Mortgage Balance
> #
> m_reg1 <- lm(assets ~ age+home+mortgage)
> summary(m_reg1)

Call:
lm(formula = assets ~ age + home + mortgage)

Residuals:
   Min     1Q Median     3Q    Max 
-37.83 -10.42   2.36  10.02  33.07 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 70.004077   4.547547  15.394  < 2e-16 ***
age          0.269996   0.087776   3.076  0.00273 ** 
home        -0.001751   0.007616  -0.230  0.81861    
mortgage     0.327100   0.038056   8.595 1.54e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.44 on 96 degrees of freedom
Multiple R-squared:  0.6029,	Adjusted R-squared:  0.5905 
F-statistic: 48.58 on 3 and 96 DF,  p-value: < 2.2e-16
```

$\widehat{\text{assets}} = 70.00 + 0.270 \text{ age } \mathbf{-0.001} \text{ homevalue} + 0.327 \text{ mortgage balance}$

Multiple Regression – R^2 and Adjusted R-square

- The R-square (R^2) of the model calculated as:
  - $SS_{Model} = 13531.4 + 3603.1 + 17610.9 = 34745.4$
  - $SS_{Total} = SS_{Model} + SS_{Residuals} = 34745.4 + 22884.8 = 57630.2$
  - $SS_{Model}/SS_{Total} = 34745.4/57630.2 = 0.6029$ tells us:
  - the three predictors together explain 60.29% of the variability in Assets
  - the remaining 39.71% is explained by all other variables not in the model (i.e., all other variables in the error term)

```
> #
> # Multiple Regression - Predict Assets using Age, Home Value and Mortgage Balance
> #
> m_reg1 <- lm(assets ~ age+home+mortgage)
> summary(m_reg1)

Call:
lm(formula = assets ~ age + home + mortgage)

Residuals:
    Min      1Q  Median      3Q     Max
-37.83  -10.42    2.36   10.02   33.07

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 70.004077   4.547547  15.394  < 2e-16 ***
age          0.269996   0.087776   3.076  0.00273 **
home        -0.001751   0.007616  -0.230  0.81861
mortgage     0.327100   0.038056   8.595 1.54e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.44 on 96 degrees of freedom
Multiple R-squared:  0.6029,	Adjusted R-squared:  0.5905
F-statistic: 48.58 on 3 and 96 DF,  p-value: < 2.2e-16

> print(anova(m_reg1))
Analysis of Variance Table

Response: assets
          Df  Sum Sq Mean Sq F value    Pr(>F)
age        1 13531.4 13531.4  56.763 2.714e-11 ***
home       1  3603.1  3603.1  15.115 0.0001863 ***
mortgage   1 17610.9 17610.9  73.876 1.540e-13 ***
Residuals 96 22884.8   238.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

$$\widehat{assets} = 70.00 + 0.270\ age\ \mathbf{-0.001}\ homevalue + 0.327\ mortgage\ \ balance$$

Multiple Regression – R^2 and Adjusted R-square

- R^2 always improves when more predictors are added to the model, *regardless of whether they are useful or meaningful.*

- A modified measure called "Adjusted R-square" is often reported that increases only if the new term improves the model more than would be expected by chance.
  - Adjusted R-square = R^2 - $(1-R^2)$ (k/n–k–1) = 59.05%

- Adjusted R-square adjusts for the number of predictors (k) in a model relative to the number of data points (n). If k is small relative to n, then the difference with the usual R^2 is small.

- In other words, Adjusted R-square tries to penalize a model where too many predictors are added just to inflate R^2. When the added predictors are not significant you will see larger differences in R^2 and Adjusted R-square, especially when k is comparable to n.

- In our case, k is 3 and n is 100, so we expect the difference between R^2 and Adjusted R-square to be small. Predictor Home Value is not significant and removing it will reduce this difference.

```
> #
> # Multiple Regression - Predict Assets using Age, Home Value and Mortgage Balance
> #
> m_reg1 <- lm(assets ~ age+home+mortgage)
> summary(m_reg1)

Call:
lm(formula = assets ~ age + home + mortgage)

Residuals:
   Min     1Q Median     3Q    Max 
-37.83 -10.42   2.36  10.02  33.07 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 70.004077   4.547547  15.394  < 2e-16 ***
age          0.269996   0.087776   3.076  0.00273 ** 
home        -0.001751   0.007616  -0.230  0.81861    
mortgage     0.327100   0.038056   8.595 1.54e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.44 on 96 degrees of freedom
Multiple R-squared:  0.6029,	Adjusted R-squared:  0.5905 
F-statistic: 48.58 on 3 and 96 DF,  p-value: < 2.2e-16

> print(anova(m_reg1))
Analysis of Variance Table

Response: assets
          Df  Sum Sq Mean Sq F value    Pr(>F)    
age        1 13531.4 13531.4  56.763 2.714e-11 ***
home       1  3603.1  3603.1  15.115 0.0001863 ***
mortgage   1 17610.9 17610.9  73.876 1.540e-13 ***
Residuals 96 22884.8   238.4                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 
```

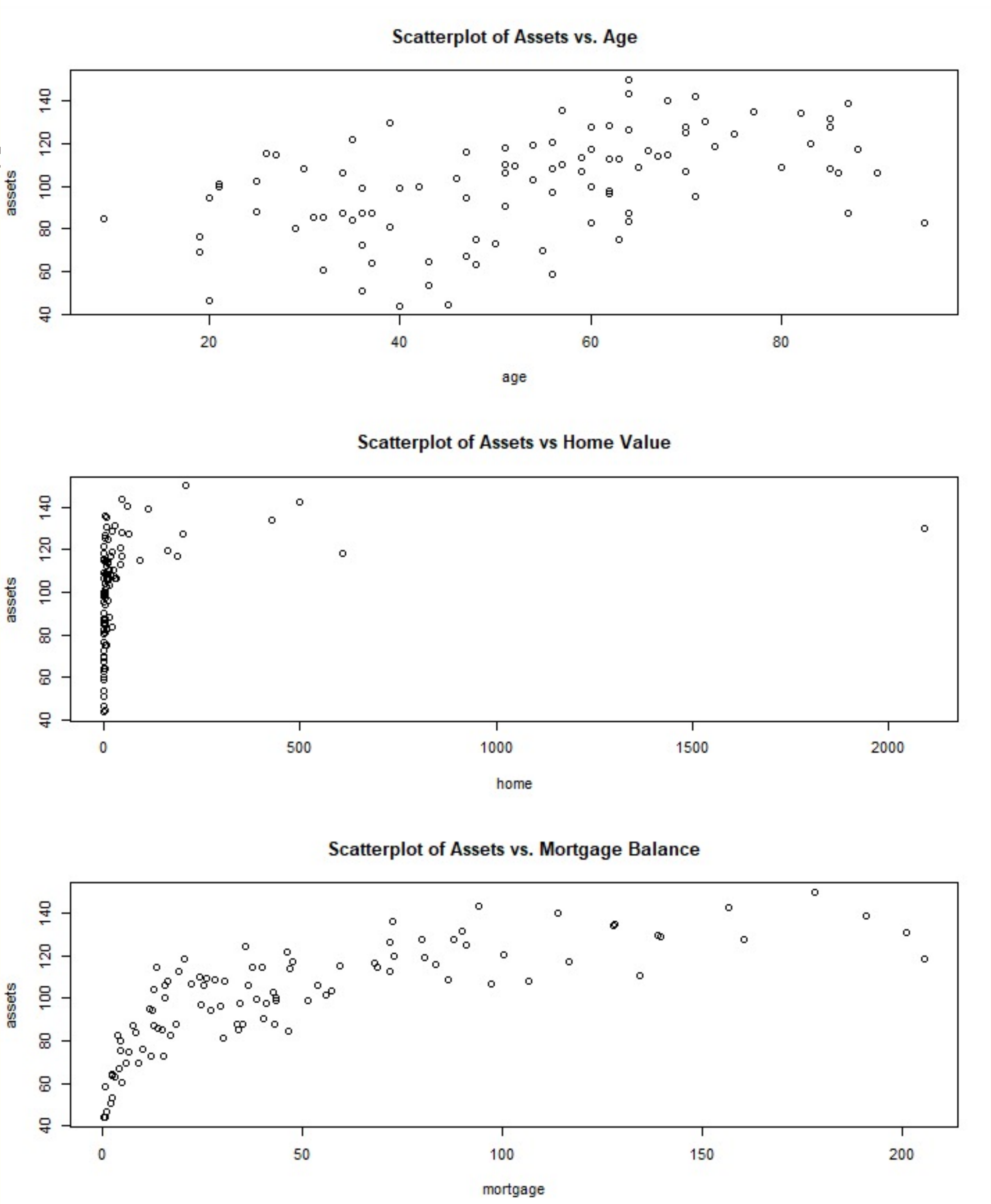
$$\widehat{assets} = 70.00 + 0.270\ age \ \mathbf{-0.001}\ homevalue + 0.327\ mortgage\ \ balance$$

Multiple Regression – Scatterplots

- It is always a good idea to plot the dependent variable against each predictor to check their relationships.
- In **R**, you can use the *par()* function to split the plot area.
  - For example, par(3,1) creates a plot area with 3 rows and 1 column, that can hold the three plots.

```
#
par(mfrow=c(3,1))
plot(age,assets, main="Scatterplot of Assets vs. Age")
plot(home,assets, main="Scatterplot of Assets vs Home Value")
plot(mortgage,assets, main="Scatterplot of Assets vs. Mortgage Balance")
#
# Simple Regression - Predict Assets using only Home Value
```

- We notice that both Home Value and Mortgage Balance have a non-linear relationship with Assets, which could affect the conclusions from our model.
- We will deal with this shortly.



Multiple Regression – Prediction

- The predicted values of assets are obtained by simply plugging in the corresponding Age, Home Value and Mortgage Balance values in the equation:
  - $\widehat{\text{assets}} = 70.00 + 0.270\ \text{age}\ \mathbf{-0.001}\ \text{homevalue} + 0.327$ mortgage balance

- To obtain predicted values in **R**:
  - *predict(model_name)* function:

- To obtain residuals in **R**:
  - *residuals(model_name)* function

```
> df$Pred_assets <- predict(m_reg1)
> df$Resid_assets <- residuals(m_reg1)
> print(head(df))
  Obs Gender Marital_Status Age Home_Value Mortgage_Balance Assets Pred_assets Resid_assets
1   1      0              1  71     498.68           156.33 142.49   139.43589    3.0541060
2   2      1              1  60      16.21            47.43 117.29   101.68977   15.6002301
3   3      1              1  82     427.71           127.70 134.15   133.16528    0.9847213
4   4      1              1  87     113.15           191.01 139.08   155.77485  -16.6948474
5   5      0              0  21       0.07            15.58 100.01    80.77008   19.2399220
6   6      0              0  31       0.17            13.85  85.72    82.90398    2.8160219
```

- To obtain the predicted Assets for a particular values of predictors in R, you create a data frame containing the values of the predictors and use it in the *predict()* function.
  - We predict Expected Assets for Age=50, Home Value = 100 (in 1000's of dollars) and Mortgage Balance = 50 (in 1000's of dollars)

```
> #
> # Predict Assets using m_reg1 - Original Model
> # Predict Assets using Age = 50, Home Value = 100 (in 1000's of dollars)
> # and Mortgage Balance = 50 (in 1000's of dollars)
> #
> df_p <- data.frame(age=50, home=100, mortgage=50)
> print(predict(m_reg1, df_p))
       1
99.68371
```

LECTURE 4B-2 – CORRELATIONS AMONG PREDICTORS AND PREDICTOR SELECTION

Multiple Regression – Refining the Model through Predictor Selection

- **<u>Original model:</u>**
- $\widehat{\text{assets}} = 70.00 + 0.270\ \text{age}\ \mathbf{-0.001}\ \text{homevalue} + 0.327$ mortgage balance
- We saw that home was not a significant predictor of assets, so we may consider dropping home and re-running the model
- However, as we saw earlier, the ***significance of a predictor may depend on what other variables are included in the model***.
- This is the ***consequence of the correlations among the predictors***.
- ***If the predictors were uncorrelated with each other, and were only correlated with the dependent variable, then adding and dropping predictors will not affect the significance of variables already in the model***.
- Further, ***in large data sets*** i.e., in data sets where sample size is very large, even meaningless predictors may be significant. Hence, statistical significance may not be the major criterion for including or excluding variables in a multiple regression model in such data sets

```
> #
> # Multiple Regression - Predict Assets using Age, Home Value and Mortgage Balance
> #
> m_reg1 <- lm(assets ~ age+home+mortgage)
> summary(m_reg1)

Call:
lm(formula = assets ~ age + home + mortgage)

Residuals:
   Min     1Q Median     3Q    Max 
-37.83 -10.42   2.36  10.02  33.07 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 70.004077   4.547547  15.394  < 2e-16 ***
age          0.269996   0.087776   3.076  0.00273 ** 
home        -0.001751   0.007616  -0.230  0.81861    
mortgage     0.327100   0.038056   8.595 1.54e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.44 on 96 degrees of freedom
Multiple R-squared:  0.6029,	Adjusted R-squared:  0.5905 
F-statistic: 48.58 on 3 and 96 DF,  p-value: < 2.2e-16
```

Multiple Regression – Choosing Predictors Using Stepwise Regression

- There are several approaches to choosing predictors, understanding that *predictors already in the model influence the statistical significance of variables being entered in the model.*
- In **R**, we will use the “olsrr” library.
- First (one time only) use :
  - **install.packages(“olsrr”)**
- In the code:
  - **library(olsrr)**
- Three functions:
  - ols_step_forward_p(*model*, penter = 0.1, details = …)
  - ols_step_backward_p(*model*, prem = 0.3, details = …)
  - ols_step_both_p(*model*, penter = 0.1, prem = 0.3, details = …)
  - (Note: details = TRUE prints details of every step; Can result in a large output)
- Specify the variables in a data frame and create a skeleton model:
  - df1 <- data.frame(assets, age, home, mortgage)
  - step_model <- lm(assets ~ . , data=df1)

```r
library(olsrr)
df1 <- data.frame(assets, age, home, mortgage)
step_model <- lm(assets ~ ., data=df1)

#ols_step_forward_p(step_model, penter=0.10, details=FALSE)
```

Multiple Regression – Forward Selection

- Starting with $\hat{y} = \hat{\propto}$ (empty model), all potential predictors are assessed and compared to a **penter** (=0.1 in our case); the variable with the **lowest p-value less than penter** enters it into the equation.

- Remaining predictors are re-evaluated given the new equation ($\hat{y} = \hat{\propto} + \hat{\beta}_{first}X_{first}$) and the next variable with the **lowest p-value less than penter** enters, etc…

- Continues until either all of the variables are entered or no other variables meet the entry criterion **penter**.

- Once variables enter the equation they **remain** (even if they subsequently have p-values greater than **penter** due to the other variables that have entered the model).

```
Final Model Output
------------------

                          Model Summary
--------------------------------------------------------------
R                       0.776       RMSE               15.364
R-Squared               0.603       Coef. Var          15.226
Adj. R-Squared          0.594       MSE               236.056
Pred R-Squared          0.574       MAE                12.175
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                               ANOVA
--------------------------------------------------------------------
                Sum of
               Squares        DF    Mean Square      F         Sig.
--------------------------------------------------------------------
Regression    34732.747        2      17366.373    73.569    0.0000
Residual      22897.427       97        236.056
Total         57630.174       99
--------------------------------------------------------------------

                                  Parameter Estimates
---------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
---------------------------------------------------------------------------------------
(Intercept)    69.940         4.517                 15.485    0.000    60.975    78.904
   mortgage     0.323         0.034        0.661     9.477    0.000     0.256     0.391
        age     0.273         0.086        0.221     3.161    0.002     0.102     0.444
---------------------------------------------------------------------------------------

                            Selection Summary
-----------------------------------------------------------------------
        Variable                  Adj.
Step    Entered     R-Square    R-Square     C(p)        AIC        RMSE
-----------------------------------------------------------------------
   1    mortgage      0.5618      0.5573    9.9484    842.9540    16.0536
   2    age           0.6027      0.5945    2.0529    835.1487    15.3641
-----------------------------------------------------------------------
```

$\widehat{assets} = 69.94 + 0.273\ \text{age} + 0.323\ \text{mortgage balance};\ R^2 = 0.603$

Multiple Regression – Backward Selection

- Starting with the full model (with all predictors) all potential predictors are assessed and compared to a **prem** (in our case = 0.3); the variable with the **highest p-value greater than prem** is removed from the equation.
- If no variables met this criterion, full model is final model.
- The process continues to remove variables that do not meet the **prem** criterion.
- *Once variables are removed, the cannot re-enter.*

```
#ols_step_backward_p(step_model, prem=0.30, details=FALSE)
```

```
Final Model Output
------------------

                            Model Summary
---------------------------------------------------------------
R                        0.776       RMSE              15.364
R-Squared                0.603       Coef. Var         15.226
Adj. R-Squared           0.594       MSE              236.056
Pred R-Squared           0.574       MAE               12.175
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
---------------------------------------------------------------------
                 Sum of
                Squares        DF    Mean Square      F         Sig.
---------------------------------------------------------------------
Regression    34732.747         2      17366.373    73.569    0.0000
Residual      22897.427        97        236.056
Total         57630.174        99
---------------------------------------------------------------------

                                  Parameter Estimates
----------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
----------------------------------------------------------------------------------------
(Intercept)    69.940         4.517                15.485    0.000     60.975    78.904
        age     0.273         0.086        0.221    3.161    0.002      0.102     0.444
   mortgage     0.323         0.034        0.661    9.477    0.000      0.256     0.391
----------------------------------------------------------------------------------------


                          Elimination Summary
------------------------------------------------------------------------
        Variable                  Adj.
Step    Removed      R-Square    R-Square     C(p)       AIC        RMSE
------------------------------------------------------------------------
   1    home          0.6027      0.5945     2.0529    835.1487    15.3641
------------------------------------------------------------------------
```

$$\widehat{\text{assets}} = 69.94 + 0.273\,\text{age} + 0.323\,\text{mortgage balance};\; R^2 = 0.603$$

Multiple Regression – Stepwise (both Forward and Backward) Selection

```
ols_step_both_p(step_model, penter = 0.10, prem = 0.30, details = FALSE)
```

- Stepwise regression is a modification of the forward selection so that after each step in which a variable was added (based on **penter**) all candidate variables in the model are checked to see if their significance has been reduced (i.e., they are above **prem**)
- If a variable with p-value > **prem** is found, it is removed from the model, but put back on the candidate list.
- A different variable from the candidate list is chosen based on **penter** to enter the model, and all variables in the model are re-evaluated based on **prem**.
- The process continues until all variables that are not in the model can no longer meet the **penter** criterion without also meeting the **prem** criterion.
- The cutoff probability for adding variables should (**penter**) be less than the cutoff probability for removing variables (**prem**) so that the procedure does not get into an infinite loop

```
Final Model Output
------------------

                           Model Summary
-----------------------------------------------------------------
R                       0.776       RMSE                  15.364
R-Squared               0.603       Coef. Var             15.226
Adj. R-Squared          0.594       MSE                  236.056
Pred R-Squared          0.574       MAE                   12.175
-----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                               ANOVA
----------------------------------------------------------------------
               Sum of
              Squares        DF    Mean Square      F          Sig.
----------------------------------------------------------------------
Regression    34732.747       2      17366.373    73.569      0.0000
Residual      22897.427      97        236.056
Total         57630.174      99
----------------------------------------------------------------------

                                  Parameter Estimates
-------------------------------------------------------------------------------------------
      model        Beta    Std. Error    Std. Beta       t        Sig      lower     upper
-------------------------------------------------------------------------------------------
(Intercept)      69.940         4.517                 15.485    0.000     60.975    78.904
   mortgage       0.323         0.034        0.661     9.477    0.000      0.256     0.391
        age       0.273         0.086        0.221     3.161    0.002      0.102     0.444
-------------------------------------------------------------------------------------------

                               Stepwise Selection Summary
-----------------------------------------------------------------------------------------
                      Added/                   Adj.
Step    Variable     Removed     R-Square    R-Square     C(p)        AIC         RMSE
-----------------------------------------------------------------------------------------
   1    mortgage     addition       0.562       0.557    9.9480    842.9540    16.0536
   2      age        addition       0.603       0.594    2.0530    835.1487    15.3641
-----------------------------------------------------------------------------------------
```

$$\widehat{\text{assets}} = 69.94 + 0.273\,\text{age} + 0.323\,\text{mortgage balance};\; R^2 = 0.603$$

Multiple Regression – Standardized Model – Comparing Predictors

- We can standardize the variables using (obs – mean)/stdev and run the standardized model:
  - $\hat{z}_{assets} = \hat{b}_{age} z_{age} + \hat{b}_{home} z_{home} + \hat{b}_{mortgage} z_{mortgage}$
  - Generally, the intercept $\hat{a}$ must be set to zero
- Standardized models help us determine which independent variable contributes most in explaining the dependent variable, based on standardized coefficients, because all the variables are unitless, and measured in terms of their standard deviations.
- One *standard deviation* increase in Mortgage Balance increases assets by 0.669 standard deviations, controlling for (or "holding constant") Age and Home Value.
- Home Mortgage is more than 3 times stronger than Age in determining Assets because $\hat{b}_{mortgage} = 0.669$ and $\hat{b}_{age} = 0.218$.
- The test for standardized coefficients requires bootstrapping since normality of the sampling distribution of the standardized coefficients is not guaranteed. *We generally do not rely on the significance of standardized coefficients, when we have the significance tests for the unstandardized coefficients.*

```
> #
> # The Standardized Multiple Regression Model
> #
> z_assets <- (assets - mean(assets))/sd(assets)
> z_age <- (age - mean(age))/sd(age)
> z_home <- (home - mean(home))/sd(home)
> z_mortgage <- (mortgage - mean(mortgage))/sd(mortgage)
> #
> z_m_reg1 <- lm(z_assets ~ z_age+z_home+z_mortgage)
> summary(z_m_reg1)

Call:
lm(formula = z_assets ~ z_age + z_home + z_mortgage)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.56783 -0.43182  0.09782  0.41515  1.37048 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.252e-17  6.399e-02   0.000  1.00000    
z_age        2.182e-01  7.094e-02   3.076  0.00273 ** 
z_home      -1.643e-02  7.146e-02  -0.230  0.81861    
z_mortgage   6.692e-01  7.786e-02   8.595 1.54e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6399 on 96 degrees of freedom
Multiple R-squared:  0.6029,	Adjusted R-squared:  0.5905 
F-statistic: 48.58 on 3 and 96 DF,  p-value: < 2.2e-16
```

LECTURE 4B-3 – MULTIPLE REGRESSION WITH CATEGORICAL PREDICTORS

Simple Regression – Categorical Predictors

- We can look at the effect of gender on assets. For now, we won't use the other variables. I am showing some of the 100 observations

Obs	Gender	Marital_Status	Age	Home_Value	Mortgage_Balance	Assets
1	0	1	71	498.68	156.33	142.49
2	1	1	60	16.21	47.43	117.29
3	1	1	82	427.71	127.7	134.15
4	1	1	87	113.15	191.01	139.08
5	0	0	21	0.07	15.58	100.01
6	0	0	31	0.17	13.85	85.72
7	0	1	86	11.58	36.61	106.2
8	0	1	37	0.08	2.49	64.02
9	0	1	56	25.2	106.63	108.16
10	0	1	62	4.92	34.49	97.62
11	0	1	95	7.01	17.14	82.68
12	0	0	34	1.04	12.89	87.33
13	0	0	64	2.79	71.69	126.67
14	0	0	64	3.57	43.18	87.73

- Population Model:
  - Assets = $\boldsymbol{\alpha} + \boldsymbol{\beta}_{\mathbf{gender}}$Gender + $\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{2})$.
  - For Femalles: Mean Assets = $\boldsymbol{\alpha} = \boldsymbol{\mu}_{\mathbf{female}}$
  - For Males: Mean assets = $\boldsymbol{\alpha} + \boldsymbol{\beta}_{\mathbf{gender}} = \boldsymbol{\mu}_{\mathbf{male}}$

- Our sample prediction model would be:
  - $\widehat{assets} = \hat{\propto} + \hat{\beta}_{gender}$Gender
  - For Females: Mean $\widehat{assets} = \hat{\propto}$
  - For Males: Mean $\widehat{assets} = \hat{\propto} + \hat{\beta}_{gender}$

- Note that Gender is a (0,1) integer variable, so we don't need to a create dummy variable.

- The null hypothesis for the regression is $\mathbf{H_0}$: $\boldsymbol{\beta}_{\mathbf{gender}} = 0$ vs the alternative hypothesis $\mathbf{H_a}$: $\boldsymbol{\beta}_{\mathbf{gender}} \neq 0$

- This is the same as saying $\mathbf{H_0}$: $(\boldsymbol{\mu}_{\mathbf{male}} - \boldsymbol{\mu}_{\mathbf{female}}) = 0$ vs the alternative hypothesis $\mathbf{H_a}$: $(\boldsymbol{\mu}_{\mathbf{male}} - \boldsymbol{\mu}_{\mathbf{female}}) \neq 0$

Category or Group	Variable Values	Population Group Mean	Estimated Mean
Female	Gender = 0	$\boldsymbol{\mu}_{\mathbf{female}} = \boldsymbol{\alpha}$	$\hat{\propto}$
Male	Gender = 1	$\boldsymbol{\mu}_{\mathbf{male}} = \boldsymbol{\alpha} + \boldsymbol{\beta}_{\mathbf{gender}}$	$\hat{\propto} + \hat{\beta}_{gender}$

Simple Regression with Categorical Predictors Vs t-test

- First, we run the simple regression model with gender as the independent variable.

- The prediction equation for expected assets is:
- $\widehat{assets}$ = 97.742 + 12.181 Gender

- Therefore, the estimated mean (or expected) Assets for Females is 97.742 and for Males it is 109.923

- The estimated *difference* in Assets between Males and Females is 12.181 with a p-value of 0.026.

- Thus, we conclude that at $\alpha = 0.05$, there is a significant difference in the mean Assets between Males and Females i.e., we reject the null hypothesis **H_0: β_{gender}**$= 0$ which is the same as **H_0:** $(\mu_{male} - \mu_{female}) = 0$ at $\alpha = 0.05$, based on our **regression model**.

```
> #
> # Simple Regression with Categorical Predictors vs t-test
> #
> catm_reg <-  lm(assets ~ gender)
> summary(catm_reg)

Call:
lm(formula = assets ~ gender)

Residuals:
    Min      1Q  Median      3Q     Max 
-53.812 -12.785   2.128  15.518  52.388 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   97.742      2.748   35.56   <2e-16 ***
gender        12.181      5.390    2.26    0.026 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.64 on 98 degrees of freedom
Multiple R-squared:  0.04954,	Adjusted R-squared:  0.03984 
F-statistic: 5.108 on 1 and 98 DF,  p-value: 0.02603

> an_catm_reg <- anova(catm_reg)
> print(an_catm_reg)
Analysis of Variance Table

Response: assets
          Df Sum Sq Mean Sq F value  Pr(>F)  
gender     1   2855 2854.91  5.1078 0.02603 *
Residuals 98  54775  558.93                  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Simple Regression with Categorical Predictors Vs t-test

- Next, we carry out an independent sample t-test of Assets for Males vs Assets for Females.

- The null hypothesis for the t-test is **H₀**: $(\mu_{male} - \mu_{female}) = 0$ vs the alternative hypothesis **Hₐ**: $(\mu_{male} - \mu_{female}) \neq 0$.

- We first create the two groups male_assets and female_assets by selecting the corresponding subsets from the original data frame.

- The t-test shows that, the estimated mean (or expected) Assets for Females is 97.742 and for Males it is 109.923

- The estimated *difference* in Assets between Males and Females is 12.181 with a p-value of 0.017.

- Thus, we conclude that at $\alpha = 0.05$, there is a significant difference in the mean Assets between Males and Females i.e., we reject the null hypothesis **H_0**: $(\mu_{male} - \mu_{female}) = 0$ at $\alpha = 0.05$, based on our **t-test.**

- The difference in p-values between regression and the t-test is due to slightly different assumptions resulting in different standard errors.

```
> #
> # Create Male and Female Groups
> #
> male_assets <- subset(df$Assets, df$Gender == 1)
> female_assets <- subset(df$Assets, df$Gender == 0)
> #
> # Perform Independent sampel two-sided t-test
> #
> ttest <- t.test(male_assets, female_assets, alternative = c("two.sided"),
+        mu = 0, paired = FALSE, conf.level = 0.95)
> print(ttest)

        Welch Two Sample t-test

data:  male_assets and female_assets
t = 2.4727, df = 52.393, p-value = 0.01669
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  2.297741 22.064857
sample estimates:
mean of x mean of y
109.92346  97.74216
```

Multiple Regression – Categorical Predictors

- We can look at the effect of both gender and maritalstatus on assets. For now, we won't use the other variables. I am showing some of the 100 observations

Obs	Gender	Marital_Status	Age	Home_Value	Mortgage_Balance	Assets
1	0	1	71	498.68	156.33	142.49
2	1	1	60	16.21	47.43	117.29
3	1	1	82	427.71	127.7	134.15
4	1	1	87	113.15	191.01	139.08
5	0	0	21	0.07	15.58	100.01
6	0	0	31	0.17	13.85	85.72
7	0	1	86	11.58	36.61	106.2
8	0	1	37	0.08	2.49	64.02
9	0	1	56	25.2	106.63	108.16
10	0	1	62	4.92	34.49	97.62
11	0	1	95	7.01	17.14	82.68
12	0	0	34	1.04	12.89	87.33
13	0	0	64	2.79	71.69	126.67
14	0	0	64	3.57	43.18	87.73

- Population Model:
  - assets = $\boldsymbol{\alpha}$ + $\boldsymbol{\beta}_{\mathbf{gender}}$gender + $\boldsymbol{\beta}_{\mathbf{marital}}$maritalstatus + $\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon}$ ~ N(0, $\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^2$).

- Our sample prediction model would be:
  - $\widehat{\text{assets}} = \hat{\alpha} + \hat{\beta}_{gender}\text{gender} + \hat{\beta}_{marital}\text{maritalstatus}$

- Note that both Gender and Marital Status are (0,1) variables, so we don't need to create dummy variables for them. So:

Category or Group	**Variable Values**	**Population Group Mean**	**Estimated Mean**
Female, Unmarried	Gender = 0, Marital = 0	$\boldsymbol{\mu}_{\mathbf{Female,Unmarried}} = \boldsymbol{\alpha}$	$\hat{\alpha}$
Female, Married	Gender = 0, Marital = 1	$\boldsymbol{\mu}_{\mathbf{Female,Married}} = \boldsymbol{\alpha} + \boldsymbol{\beta}_{\mathbf{marital}}$	$\hat{\alpha} + \hat{\beta}_{marital}$
Male, Unmarried	Gender = 1, Marital = 0	$\boldsymbol{\mu}_{\mathbf{Male,Unmarried}} = \boldsymbol{\alpha} + \boldsymbol{\beta}_{\mathbf{gender}}$	$\hat{\alpha} + \hat{\beta}_{gender}$

Multiple Regression – Categorical Predictors

```
> #
> # Multiple Regression with Categorical Variables
> #
> catm_reg1 <- lm(assets ~ gender+marital)
> summary(catm_reg1)

Call:
lm(formula = assets ~ gender + marital)

Residuals:
   Min     1Q Median     3Q    Max 
-54.38 -13.30   1.63  16.30  51.83 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   96.413      4.870  19.797   <2e-16 ***
gender        11.764      5.559   2.116   0.0369 *  
marital        1.892      5.709   0.331   0.7410    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.75 on 97 degrees of freedom
Multiple R-squared:  0.05061,	Adjusted R-squared:  0.03104 
F-statistic: 2.586 on 2 and 97 DF,  p-value: 0.08054
```

Category or Group	**Variable Values**	**Estimated Mean**	**Value**
Female, Unmarried	Gender = 0, Marital = 0	$\hat{\propto}$	96.41
Female, Married	Gender = 0, Marital = 1	$\hat{\propto} + \hat{\beta}_{marital}$	98.35
Male, Unmarried	Gender = 1, Marital = 0	$\hat{\propto} + \hat{\beta}_{gender}$	108.17

Multiple Regression – Categorical Predictors

- Population Model:
  - assets = $\boldsymbol{\alpha}$ + $\boldsymbol{\beta}_{\mathbf{gender}}$gender + $\boldsymbol{\beta}_{\mathbf{marital}}$maritalstatus + $\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\sigma}_{\varepsilon}^2)$.
- Our sample prediction model would be, (for mean or expected or predicted assets):
  - $\widehat{\text{assets}} = \hat{\alpha} + \hat{\beta}_{\text{gender}}\text{gender} + \hat{\beta}_{\text{marital}}\text{maritalstatus}$
- Remember that in multiple regression,
  - the null hypothesis for the F-test is: $\mathbf{H_0}$: $\boldsymbol{\beta}_{\mathbf{gender}} = \boldsymbol{\beta}_{\mathbf{marital}} = 0$
  - with alternate hypothesis $\mathbf{H_a}$: At least one of $\boldsymbol{\beta}_{\mathbf{gender}}$, $\boldsymbol{\beta}_{\mathbf{marital}} \neq 0$
- This translates therefore (based on the table showing population means for the groups):
  - H_0: $\mu_{\text{Female,Unmarried}} = \mu_{\text{Female,Married}} = \mu_{\text{Male,Unmarried}} = \mu_{\text{Male,Married}} =$ **(intercept α)**
  - H_a: at least one of $\mu_{\text{Female,Unmarried}}$, $\mu_{\text{Female,Married}}$, $\mu_{\text{Male,Unmarried}}$, $\mu_{\text{Male,Married}}$ is different from the others.
- For our dataset we conclude, based on the p-value of 0.0854 for the F-test that **there is no significant difference** among the mean Assets of the four groups at $\alpha = 0.05$

```
> #
> # Multiple Regression with Categorical Variables
> #
> catm_reg1 <- lm(assets ~ gender+marital)
> summary(catm_reg1)

Call:
lm(formula = assets ~ gender + marital)

Residuals:
   Min     1Q Median     3Q    Max 
-54.38 -13.30   1.63  16.30  51.83 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   96.413      4.870  19.797   <2e-16 ***
gender        11.764      5.559   2.116   0.0369 *  
marital        1.892      5.709   0.331   0.7410    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.75 on 97 degrees of freedom
Multiple R-squared:  0.05061,	Adjusted R-squared:  0.03104 
F-statistic: 2.586 on 2 and 97 DF,  p-value: 0.08054
```

Multiple Regression – Categorical Predictors

- However, when we look at the test for the individual coefficients, we notice that the slope of the Gender variable is significant at $\alpha = 0.05$
- This says that $\mathbf{H_0}$: $\boldsymbol{\beta}_{\mathbf{gender}} = 0$ is rejected at $\alpha = 0.05$.
- This implies:
  - $\mu_{Male,Unmarried}\ (= \boldsymbol{\alpha} + \boldsymbol{\beta}_{gender}) - \mu_{Female,Unmarried}\ (= \boldsymbol{\alpha}) = \boldsymbol{\beta}_{gender} \neq 0$
  - $\mu_{Male,Married}\ (= \boldsymbol{\alpha} + \boldsymbol{\beta}_{gender} + \boldsymbol{\beta}_{marital}) - \mu_{Female,Married}\ (= \boldsymbol{\alpha} + \boldsymbol{\beta}_{marital}) = \boldsymbol{\beta}_{gender} \neq 0$
- i.e., mean assets of Males (married or unmarried) are significantly different from mean assets of (married or unmarried) females
- This result is an anomaly (false result) caused by the fact that there are only 2 single, males in the data set resulting in inadequate sample size for that group.

```
> #
> # Multiple Regression with Categorical Variables
> #
> catm_reg1 <- lm(assets ~ gender+marital)
> summary(catm_reg1)

Call:
lm(formula = assets ~ gender + marital)

Residuals:
   Min     1Q Median     3Q    Max
-54.38 -13.30   1.63  16.30  51.83

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   96.413      4.870  19.797   <2e-16 ***
gender        11.764      5.559   2.116   0.0369 *
marital        1.892      5.709   0.331   0.7410
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.75 on 97 degrees of freedom
Multiple R-squared:  0.05061,   Adjusted R-squared:  0.03104
F-statistic: 2.586 on 2 and 97 DF,  p-value: 0.08054
```

```
> df2 <- data.frame(gender, marital)
> table(df2)
      marital
gender  0  1
     0 22 52
     1  2 24
```

LECTURE 4B-4 – MULTIPLE REGRESSION WITH CATEGORICAL PREDICTORS

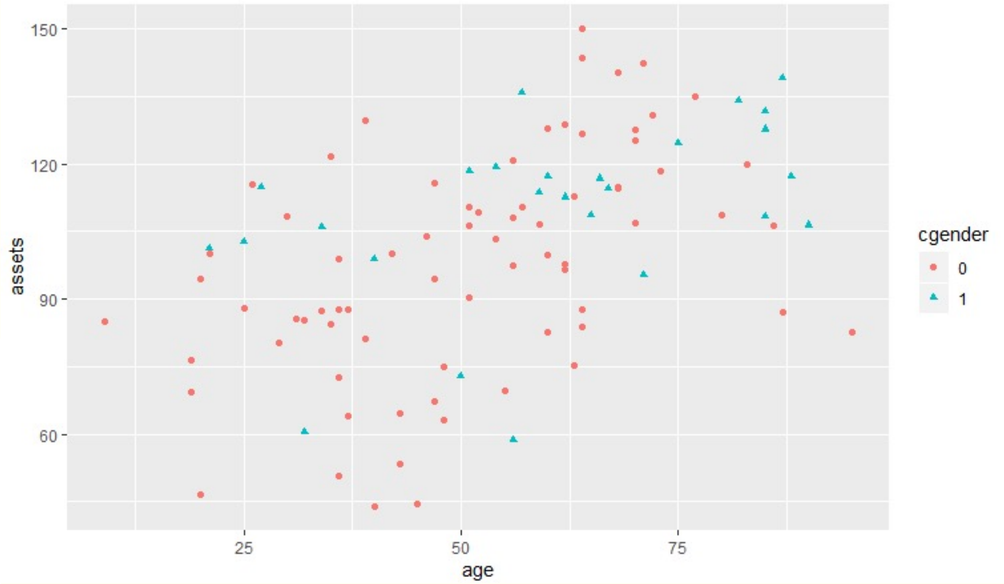
Multiple Regression – Categorical and Continuous Predictors

- We will fit a regression for Assets that includes both categorical (Gender) and Continuous (Age) Variables
  - Population Model:
    - Assets = $\boldsymbol{\alpha} + \boldsymbol{\beta}_{\mathbf{gender}}\text{gender} + \boldsymbol{\beta}_{\mathbf{age}}\text{age} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\sigma}_{\varepsilon}^{2})$.
  - Sample Prediction Model
    - $\widehat{assets} = \widehat{\propto} + \widehat{\beta}_{gender}\text{gender} + \widehat{\beta}_{age}\text{age}$

- There are two regression equations:
  - (For Females) $\widehat{assets} = \widehat{\propto} + \widehat{\beta}_{age}\text{age}$
  - (For Males) $\widehat{assets} = (\widehat{\propto} + \widehat{\beta}_{gender}) + \widehat{\beta}_{age}\text{age}$

Multiple Regression – Categorical and Continuous Predictors

- Scatterplots of assets vs age and mortgage by gender group.
  - We use the **ggplot2** library for this.




```
# Multiple Regression with Categorical and Continuous Predictors
#
# First let us plot assets against age for the gender group
#
# Create cgender as a factor from the integer variable gender, to be used in ggplot
#
cgender <- factor(gender)
#
# Use the ggplot2 library
#
library(ggplot2)
#
# ggplot() requires data in the data frame; we will use our original data frame df
#
# aes in ggplot provides the axes. Setting color and shape to the cgender variable allows
# the identification of the gender group in the plot
# The geom_point() function allows the color, size and shape of the points to be set.
#
ggplot(df, aes(x=age, y=assets, shape=cgender, color=cgender)) + geom_point()
#
```

Multiple Regression – Categorical and Continuous Predictors

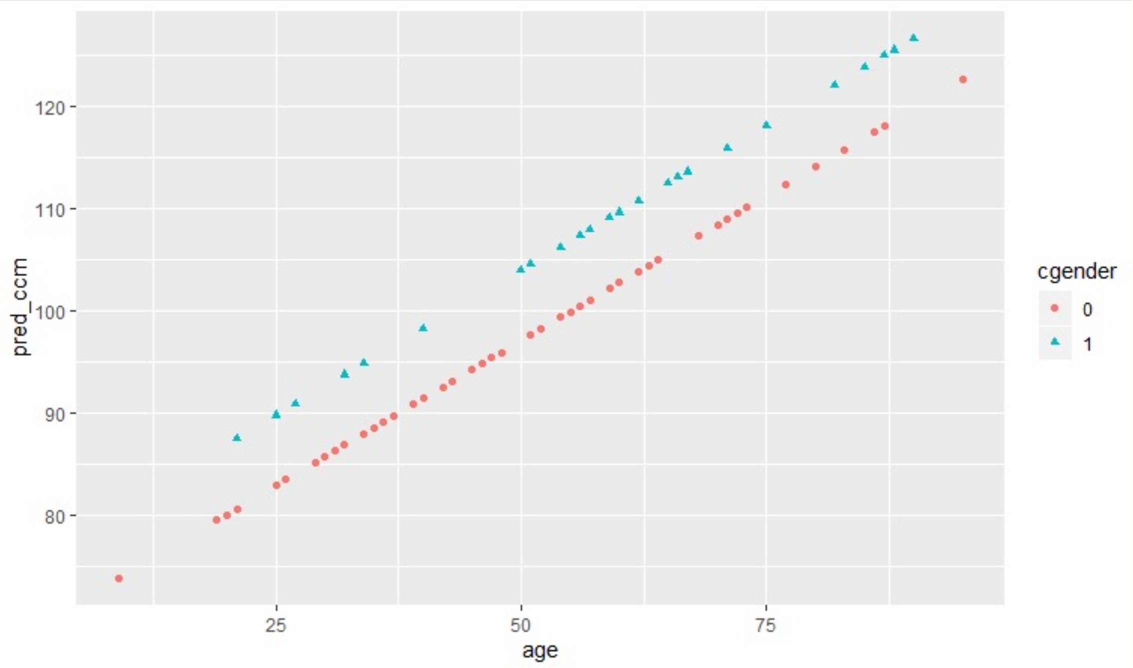
- We will fit a regression for Assets that includes both categorical (Gender) and Continuous (Age) Variables
  - Population Model:
    - Assets = $\boldsymbol{\alpha}$ + $\boldsymbol{\beta}_{\mathbf{gender}}$gender + $\boldsymbol{\beta}_{\mathbf{age}}$age + $\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\sigma}_{\varepsilon}^{2})$.
  - Sample Prediction Model
    - $\widehat{\text{assets}} = \hat{\alpha} + \hat{\beta}_{\text{gender}}\text{gender} + \hat{\beta}_{\text{age}}\text{age}$
- There are two regression equations:
  - (For Females) $\widehat{\text{assets}} = 68.742 + 0.567$ age
  - (For Males) $\widehat{\text{assets}} = (68.742 + 6.876) + 0.567$ age
- You can then see, that the two equations differ only in the intercept by 6.876, with Males recording a higher mean asset value of 3.727 (thousands of dollars)
- The model explains 25% of the variability in assets based on R^2;.
- The overall model fit, based on ANOVA F-test is significant with a p-value = 0.000 at $\alpha = 0.05$
- Gender is not a significant predictor (controlling for age, with a p-value of 0.166) but age (controlling for gender) is a significant at $\alpha = 0.05$ with a p-value of 0.000.

```
> # Now develop the actual multiple regression model with genderand age
> #
> ccm_reg1 <- lm(assets ~ gender+age)
> summary(ccm_reg1)

Call:
lm(formula = assets ~ gender + age)

Residuals:
    Min      1Q  Median      3Q     Max
-49.793 -10.468   3.858  13.022  45.121

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  68.7422     6.2038  11.081  < 2e-16 ***
gender        6.8757     4.9245   1.396    0.166
age           0.5667     0.1113   5.090 1.76e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.11 on 97 degrees of freedom
Multiple R-squared:  0.2499,    Adjusted R-squared:  0.2344
F-statistic: 16.16 on 2 and 97 DF,  p-value: 8.791e-07

> an_ccm_reg1 <- anova(ccm_reg1)
> print(an_ccm_reg1)
Analysis of Variance Table

Response: assets
          Df Sum Sq Mean Sq F value    Pr(>F)
gender     1   2855  2854.9  6.4059   0.01298 *
age        1  11545 11545.3 25.9054 1.756e-06 ***
Residuals 97  43230   445.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple Regression – Categorical and Continuous Predictors

- There are two regression equations:
  - (For Females) $\widehat{assets}$ = 68.742 + 0.567 age
  - (For Males) $\widehat{assets}$ = (68.742 + 6.876) + 0.567 age
- Plotting the Prediction lines:

```
#
# Plot of predicted assets for each gender, by age
#
pred_ccm <- predict(ccm_reg1)
ggplot(df, aes(x=age, y=pred_ccm, shape=cgender, color=cgender)) + geom_point()
#
```



Multiple Regression – Categorical and Continuous Predictors

- We will fit a regression for Assets that includes both Categorical (Gender, Marital Status) and Continuous (Age, Mortgage Balance) Variables

- Population Model:
  - Assets = $\boldsymbol{\alpha}$ + $\boldsymbol{\beta}_{\mathbf{gender}}$gender + $\boldsymbol{\beta}_{\mathbf{marital}}$maritalstatus + $\boldsymbol{\beta}_{\mathbf{age}}$age + $\boldsymbol{\beta}_{\mathbf{mort}}$mortgagebalance + $\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon}$ ~ N($\mathbf{0}$, $\boldsymbol{\sigma}_{\varepsilon}^{2}$).

- Sample Prediction Model
  - $\widehat{assets} = \hat{\alpha} + \hat{\beta}_{gender}$gender + $\hat{\beta}_{marital}$maritalstatus + $\hat{\beta}_{age}$age + $\hat{\beta}_{mort}$mortgagebalance
  - $\widehat{assets}$ = 71.11 + 4.039gender + −1.717maritalstatus + 0.257age + 0.322mortgagebalance

- There are four regression equations:
  - (For Females, Unmarried) $\widehat{assets}$ = 71.11 + 0.257age + 0.322mortgagebalance
  - (For Females, Married) $\widehat{assets}$ = (71.11 – 1.717) + 0.257age + 0.322mortgagebalance
  - (For Males, Unmarried) $\widehat{assets}$ = (71.11 + 4.04) + 0.257age + 0.322mortgagebalance
  - (For Males, Married) $\widehat{assets}$ = (71.11 + 4.04 – 1.717) + 0.257age + 0.322mortgagebalance

- You can then see, that the four equations differ only in the intercepts reflecting differences in the mean asset value for each group

```
> # Multiple Regression Model for assets with
> # gender and marital (categorical) and
> # age and mortgage (continuous)
> #
> m_reg3 <- lm(assets ~gender+marital+age+mortgage)
> summary(m_reg3)

Call:
lm(formula = assets ~ gender + marital + age + mortgage)

Residuals:
    Min      1Q  Median      3Q     Max 
-38.377  -9.800   2.679   9.003  30.043 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 71.11203    5.22645  13.606  < 2e-16 ***
gender       4.03949    3.69775   1.092  0.27741    
marital     -1.71619    3.72591  -0.461  0.64613    
age          0.25754    0.08786   2.931  0.00423 ** 
mortgage     0.32154    0.03457   9.301 5.16e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.42 on 95 degrees of freedom
Multiple R-squared:  0.6078,	Adjusted R-squared:  0.5913 
F-statistic: 36.81 on 4 and 95 DF,  p-value: < 2.2e-16
```

```
# Predict the assets for an Unmarried Female,
# Aged 50 with a Mortgage Balance of 100 (,in 1000's of dollars)
x_vals <- data.frame(gender=0,marital=0,age=50,mortgage=100)
print(predict(m_reg3, x_vals))
```

```
> print(predict(m_reg3, x_vals))
       1 
116.1429 
```

Multiple Regression – Categorical and Continuous Predictors

- The model explains 60.78% of the variability in Assets based on R^2
- The overall model fit, based on ANOVA F-test is significant with a p-value = 0.000 at $\alpha = 0.05$
- Gender is not a significant predictor (controlling for Age, Marital and Mortgage), Marital is not a significant predictor (controlling for Age, Gender and Mortgage), but both Age (controlling for Gender, Marital and Mortgage) and Mortgage (controlling for Age, Marital and Gender) are significant at $\alpha = 0.05$

```
> # Multiple Regression Model for assets with
> # gender and marital (categorical) and
> # age and mortgage (continuous)
> #
> m_reg3 <- lm(assets ~gender+marital+age+mortgage)
> summary(m_reg3)

Call:
lm(formula = assets ~ gender + marital + age + mortgage)

Residuals:
    Min      1Q  Median      3Q     Max 
-38.377  -9.800   2.679   9.003  30.043 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 71.11203    5.22645  13.606  < 2e-16 ***
gender       4.03949    3.69775   1.092  0.27741    
marital     -1.71619    3.72591  -0.461  0.64613    
age          0.25754    0.08786   2.931  0.00423 ** 
mortgage     0.32154    0.03457   9.301 5.16e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.42 on 95 degrees of freedom
Multiple R-squared:  0.6078,	Adjusted R-squared:  0.5913 
F-statistic: 36.81 on 4 and 95 DF,  p-value: < 2.2e-16

> an_m_reg3 <- anova(m_reg3)
> print(an_m_reg3)
Analysis of Variance Table

Response: assets
          Df  Sum Sq Mean Sq F value    Pr(>F)    
gender     1  2854.9  2854.9 12.0005 0.0007996 ***
marital    1    62.0    62.0  0.2604 0.6110060    
age        1 11533.2 11533.2 48.4794 4.301e-10 ***
mortgage   1 20579.7 20579.7 86.5064 5.159e-15 ***
Residuals 95 22600.4   237.9                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 
```

Predictions and Confidence Intervals for Expected Value of Y

- We can generate predictions for expected value of Y (Assets) in the population from the regression equation, by supplying the values for the predictors (i.e., for a given X-vector). This is also the point estimate for the expected value of Y in the population, for that given X-vector.

```
> #
> # Predict the Expected Value of Assets for an Unmarried Female,
> # Aged 50 with a Mortgage Balance of 100 (,in 1000's of dollars).
> #
> # This is a point estimate for the Expected Value of Assets in the Population
> # for the given X-vector.
> x_vals <- data.frame(gender=0,marital=0,age=50,mortgage=100)
> print(predict(m_reg3, x_vals))
       1
116.1429
```

- We can also generate a ***confidence interval for the Expected value of Y*** (in the population) for a given value of the x-vector.

```
> # 95% Confidence Interval for the Expected Value of Y in the population, for the above values of X
> #
> print(predict(m_reg3, x_vals, interval="confidence", level=0.95))
       fit      lwr      upr
1 116.1429 108.5175 123.7683
```

- The interval tells us that, for this specific value of the x-vector, for samples of that size (100), if we keep constructing 95% confidence intervals, then roughly 95% of the confidence intervals will contain the corresponding ***expected value of y* (in the population)**.

Prediction Interval for Individual Actual Value of Y

- We can also generate a ***prediction interval for the actual value of Y*** (in the population) for a given value of the x-vector.

```
> #
> # 95% Prediction Interval for the actual value of Y in the population, for above values of X
> #
> print(predict(m_reg3, x_vals, interval="predict", level=0.95))
       fit     lwr      upr
1 116.1429 84.5873 147.6986
>
```

- The interval tells us that, for this specific value of the x-vector, for samples of that size (100), if we keep constructing 95% prediction intervals, then roughly 95% of the prediction intervals will contain the corresponding ***actual value of y***.

- Because, the prediction interval deals with individual values of Y, it will be wider than the confidence interval for the Expected value (or average of Y values), for the same level of confidence.

Prediction and Confidence Intervals for the all X values

- Showing only first 6 records.

```
> #
> # Create a data frame with gender, marital, age, mortagage and assets -- df2
> # Confidence and Prediction Intervals for the whole data set - conf_df
> # Create data frames to store the confidence and prediction intervals - pred_df
> # Then merge the specific columns from conf_df and pred_df with df2 to create data frame df3.
> #
> df2 <- data.frame(gender, marital, age, mortgage, assets)
> conf_df <- data.frame(predict(m_reg3, df2, interval="confidence", level=0.05))
> pred_df <- data.frame(predict(m_reg3, df2, interval="prediction", level=0.05))
> df3 <- data.frame(df2, conf_df$fit, conf_df$lwr, conf_df$upr, pred_df$lwr, pred_df$upr)
> print(head(df3))
  gender marital age mortgage assets conf_df.fit conf_df.lwr conf_df.upr pred_df.lwr pred_df.upr
1      0       1  71   156.33 142.49   137.94737   137.69295   138.20179   136.94481   138.94994
2      1       1  60    47.43 117.29   104.13839   103.94361   104.33317   103.14928   105.12750
3      1       1  82   127.70 134.15   135.61418   135.37602   135.85234   134.61562   136.61274
4      1       1  87   191.01 139.08   157.25848   156.93879   157.57816   156.23739   158.27956
5      0       0  21    15.58 100.01    81.52997    81.27862    81.78132    80.52818    82.53176
6      0       0  31    13.85  85.72    83.54913    83.32541    83.77284    82.55391    84.54434
```

Important Things to Remember about Regression

- Pairwise correlation is limited in value and can be misleading. It is uncontrolled (zero-order) correlation.
- Partial correlation controls for the correlation of the pair of variables with other variables (first-order is one other variable, second-order is two other variables etc.)
- Regression makes use of correlations in a way that permits predictions using linear models.
- In a linear regression model it is important that you understand:
  - Predictors may be correlated with each other (as well as the dependent variable).
  - Predictors in the model determine R^2 – the percentage of variability explained in the dependent variable.
  - Predictors excluded from the model are in the noise term (error term)
- Regression betas in multiple regression behave somewhat similarly to partial correlation, by controlling for the inter-correlation of a predictor with other predictors.

Important Things to Remember about Regression

- The statistical significance of an individual predictor in the model depends on what other predictors are in the model, because of the correlations of that predictor with other predictors
- The beta coefficients are always interpreted as the change in the dependent variable, for a unit change in a predictor, taking into account all the other predictors in the model (i.e., controlling for its correlation with other predictors). That is, its value is affected by what other predictors are in the model, due to inter-correlation of predictors.
- If a predictor is not correlated (uncorrelated) with other predictors in the model, its significance and beta coefficient is not changed by the presence of the other predictors.
- The sample regression model is just that; it is a model-based estimate of the population model, and the sample model will change from sample to sample.
- Very large sample sizes often result in many or all predictors being significant, regardless of whether they are meaningful predictors or not.
- The hypothesis test of population betas using the sample betas depends on the assumption of normality of the noise or error term.