LECTURE 4 - CORRELATION & REGRESSION – PART C

Testing Regression Assumptions

Testing Regression Assumptions

- 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 $\boldsymbol{e} = (y - \hat{y})$
- In order for our conclusions about the regression model (that we fit) to be valid, we need to check four main assumptions:
  - *Linearity/Nonlinearity between predictors and dependent variable*
  - *Normality of Residuals*
  - *Homoscedasticity* (constant variance of residuals across X) or the opposite, *heteroscedasticity*
  - *Statistical Independence of residuals* (relevant in time-series data)
- In addition, other problems (have to be considered and fixed) such as:
  - *Multicollinearity* of predictors (excessive correlations among predictors)
  - *Missing Data*, especially in large secondary datasets in analytics)
- Failure to address these problems could result in invalid conclusions from the regression analysis.

http://people.duke.edu/~rnau/testing.htm

LINEARITY

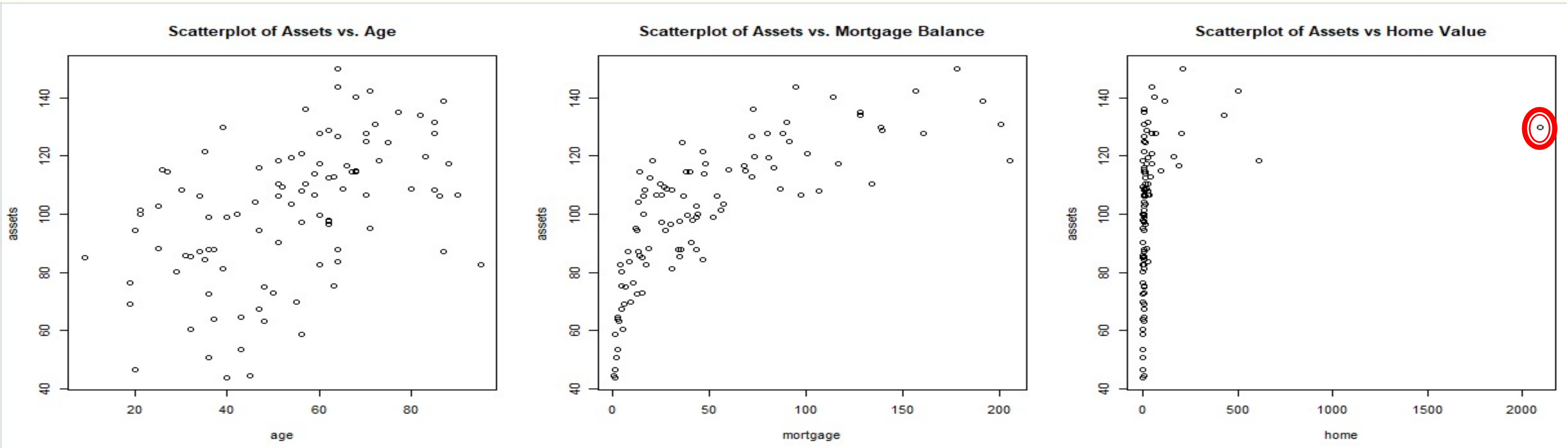
Lecture 4c-Part 1

Linearity/Nonlinearity of Relationships between Predictors and Dependent Variable - **AssumpReg.R**

- You can check whether the relationships between each predictor and the dependent variable (Y) are linear using sc

```
par(mfrow=c(1,3))
plot(age,assets, main="Scatterplot of Assets vs. Age")
plot(mortgage,assets, main="Scatterplot of Assets vs. Mortgage Balance")
plot(home,assets, main="Scatterplot of Assets vs Home Value")
```

- We are looking for:
  - Large deviations from linearity
  - Heteroscedasticity (pattern of spread)
  - Outliers

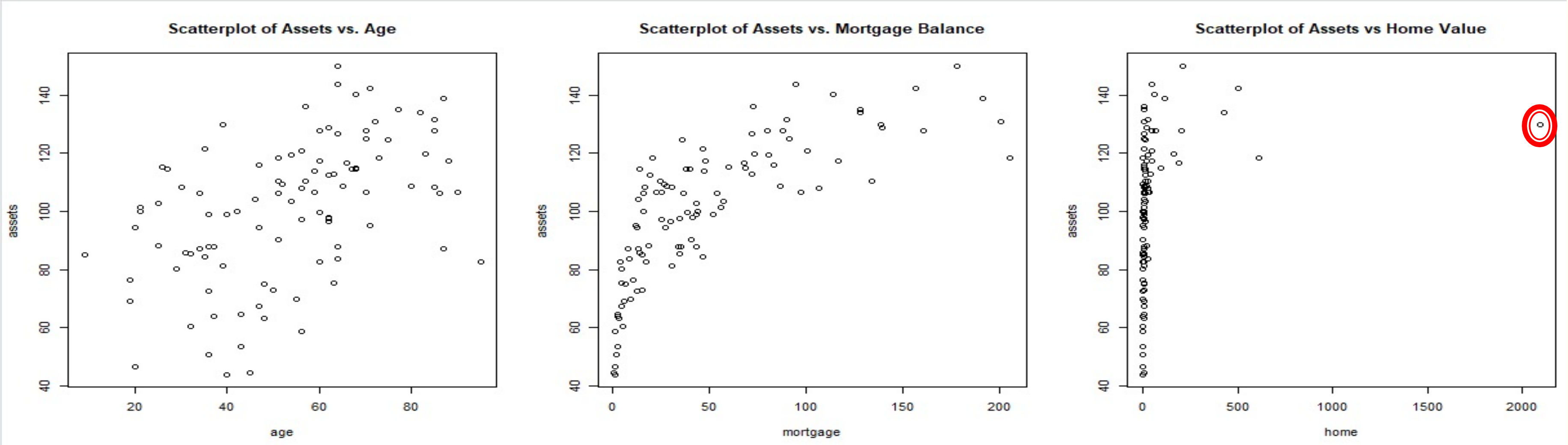


Checking plots of Y vs X – **AssumpReg.R**

- **Assets vs Age**
  - There seems to be a linear trend in Assets as Age increases. There appears to be less spread in Assets for larger values of Age. No observation seems to be an outlier.

- **Assets vs Mortgage Balance**
  - Mortgage Balance has a distinct nonlinear relationship with Assets, with distinctly less spread in Assets at the low end of Mortgage, with no outliers

- **Assets vs Home Value**
  - Home Value has a distinct nonlinear relationship with Assets, with distinctly less spread in Assets at the low end of Home Value. Further, There is one value of Home which appears to be “extreme”



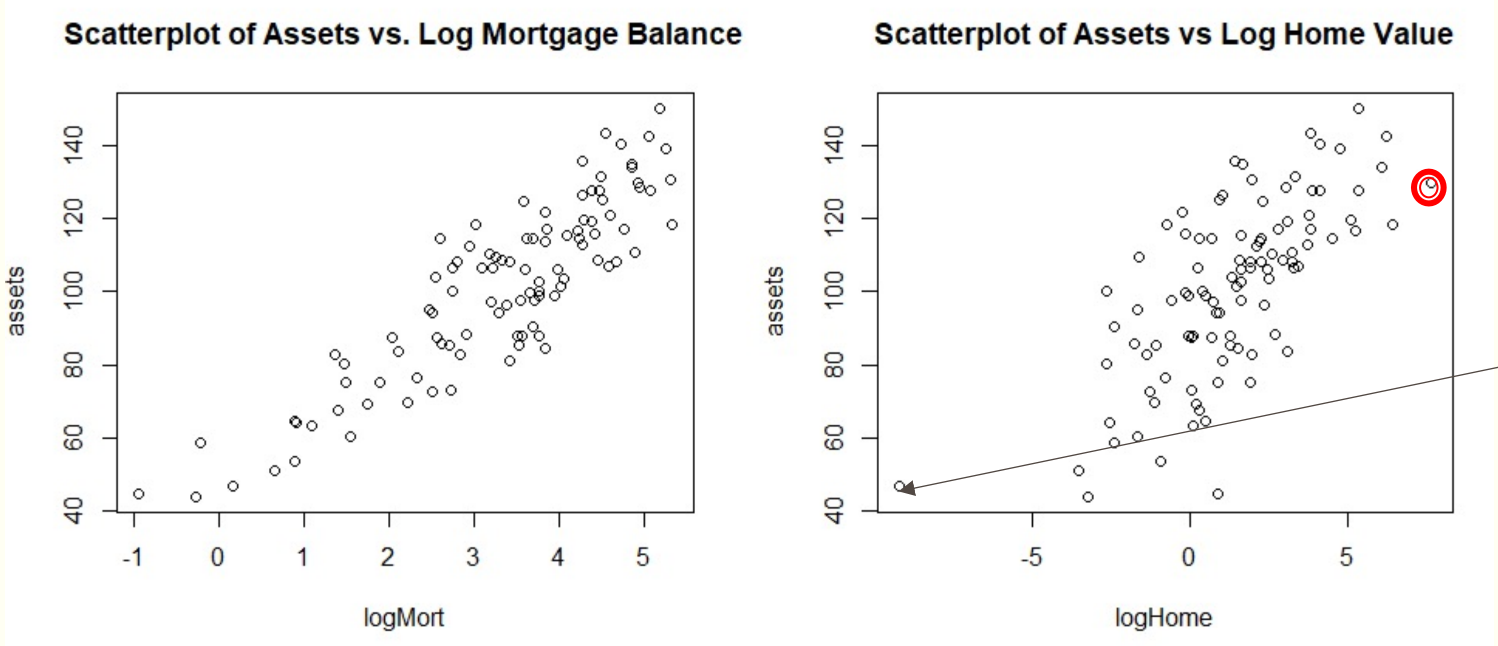
Transforming to Linearity

```
#
logMort <- log(mortgage)
logHome <- log(home + 0.0001)
#
par(mfrow=c(1,2))
plot(logMort,assets, main="Scatterplot of Assets vs. Log Mortgage Balance")
plot(logHome,assets, main="Scatterplot of Assets vs Log Home Value")
```

- We consider transforming Mortgage and Home using *log* functions. The *logic for a log transformation* is that an examination of the data shows that for large increases in Home and Mortgage on the x-axis, Assets on the Y-axis remains relatively flat after an initial sharp increase. The only caveat is that the values of the variable before log transformation have to be strictly positive.

- Because one of the values of Homevalue was 0, I added 0.001 to every value before taking the log. i.e., **logHome = log(home + 0.0001)**

- The plots of ln(Mortgage) and ln(Home) vs. Assets shows a substantially better linear relationship. Even the “extreme” value of Home from the previous overhead, no longer appears extreme.



This value corresponds to homevalue = 0

Checking the Correlation Matrices

- We can check the correlation matrices before and after transformations

- We can see clear improvements in correlations after the transformations, with logHome and logMort displaying better correlations with each other and with assets. But remember, these are uncontrolled correlations.

- The best way to see if transformations helped is to fit regression models before and after transformations.

- **mod1**: $\widehat{assets} = \hat{\alpha} + \hat{\beta}_{age}age + \hat{\beta}_{mort}mortgage$

- **mod2**: $\widehat{assets} = \hat{\alpha} + \hat{\beta}_{age}age + \hat{\beta}_{logmort}logMort + \hat{\beta}_{loghome}logHome$

```
> # Checking correlations before and after transformations
> #
> M1 <- cbind(assets, age, home, mortgage)
> print(cor(M1))
            assets        age       home  mortgage
assets   1.0000000 0.48455816 0.26937984 0.7495006
age      0.4845582 1.00000000 0.04032963 0.3989889
home     0.2693798 0.04032963 1.00000000 0.4139189
mortgage 0.7495006 0.39898894 0.41391890 1.0000000
> M2 <- cbind(assets, age, logHome, logMort)
> print(cor(M2))
           assets       age   logHome   logMort
assets  1.0000000 0.4845582 0.7056161 0.8843015
age     0.4845582 1.0000000 0.5027122 0.3744746
logHome 0.7056161 0.5027122 1.0000000 0.7093682
logMort 0.8843015 0.3744746 0.7093682 1.0000000
```

Comparing the Regression Models

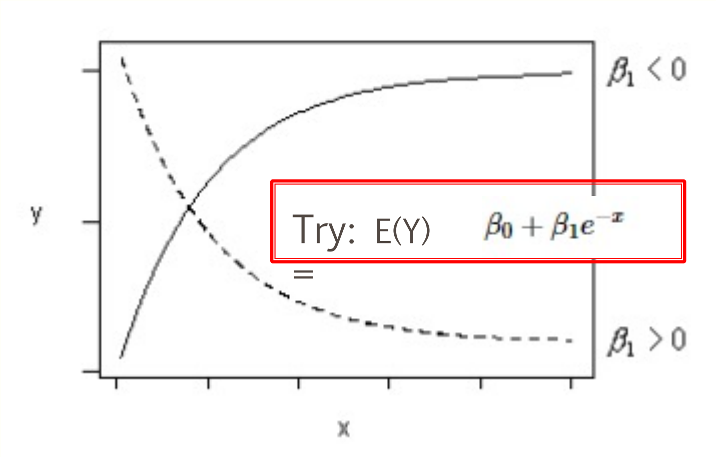
mod1: $\widehat{\text{assets}} = 69.94 + 0.273\text{age} + 0.323\text{mortgage}$
mod2: $\widehat{\text{assets}} = 44.421 + 0.192\text{age} + 13.695\text{logMort} + 0.804\text{logHome}$

- Model mod1 has the untransformed Home Value and Mortgage. Home Value is not a significant predictor. The R^2 is 0.6027
- Model mod2 has the transformed logHome and logMort predictors. logHome is not a significant predictor. The has R^2 improved to 0.8124.
- We can now proceed with checking other assumptions.
- For *illustration purposes*, we will retain logHome in Mod2, even if it is not significant.
- We will perform residual and other diagnosis for each model.

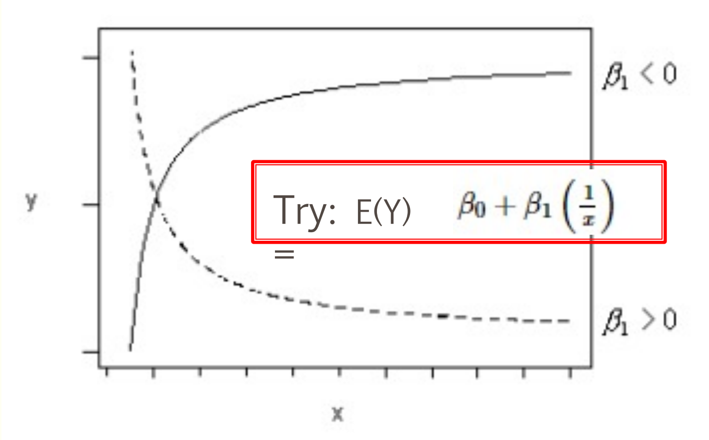
```
> mod1 <- lm(assets ~ age+mortgage)
> summary(mod1)

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

Residuals:
    Min      1Q  Median      3Q     Max 
-37.901 -10.397   2.537   9.932  33.098 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 69.93952    4.51666  15.485  < 2e-16 ***
age          0.27301    0.08636   3.161   0.0021 ** 
mortgage     0.32330    0.03411   9.477 1.81e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.36 on 97 degrees of freedom
Multiple R-squared:  0.6027,	Adjusted R-squared:  0.5945 
F-statistic: 73.57 on 2 and 97 DF,  p-value: < 2.2e-16

> mod2 <- lm(assets ~ age+logHome+logMort)
> summary(mod2)

Call:
lm(formula = assets ~ age + logHome + logMort)

Residuals:
     Min       1Q   Median       3Q      Max 
-21.5862  -6.7227  -0.5866   6.8165  29.3011 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 44.42132    4.47059   9.936  < 2e-16 ***
age          0.19230    0.06331   3.038  0.00307 ** 
logHome      0.80373    0.64795   1.240  0.21784    
logMort     13.69447    1.12055  12.221  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.61 on 96 degrees of freedom
Multiple R-squared:  0.8124,	Adjusted R-squared:  0.8065 
F-statistic: 138.5 on 3 and 96 DF,  p-value: < 2.2e-16

> |
```

Common Transformations

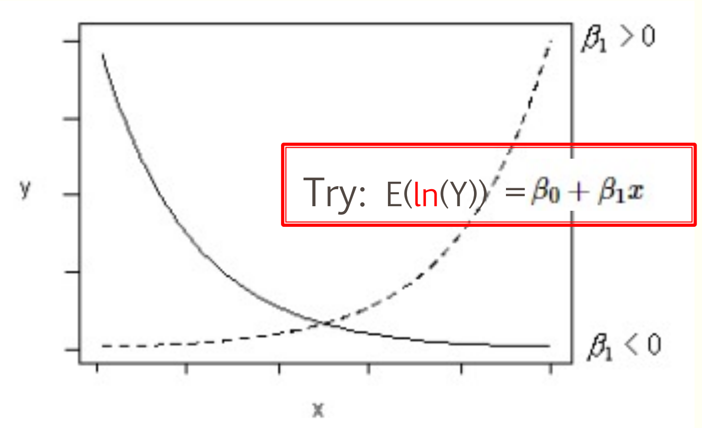
https://newonlinecourses.science.psu.edu/stat462/node/155/



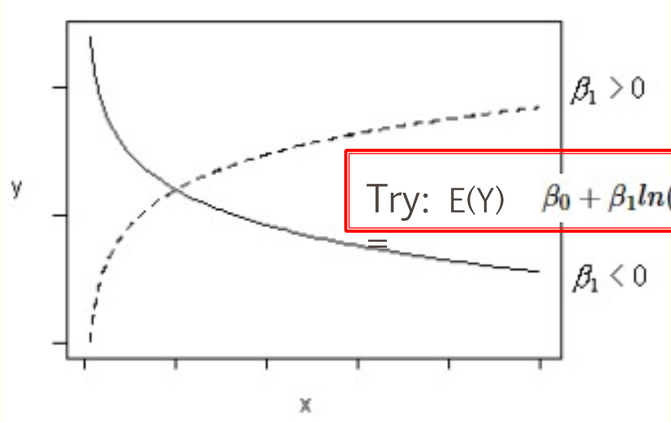
Try: E(Y) = $\beta_0 + \beta_1 e^{-x}$



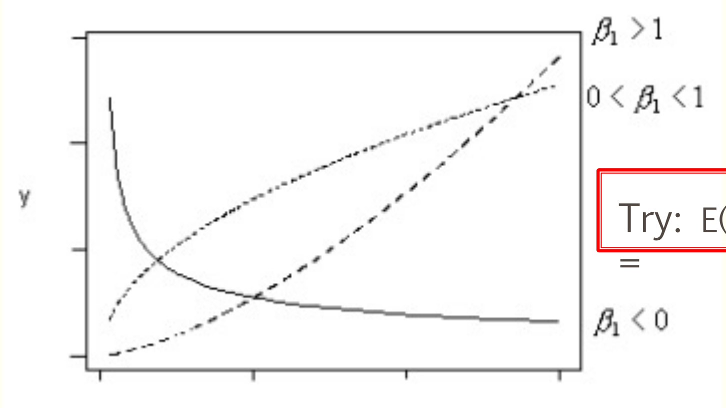
Try: E(Y) = $\beta_0 + \beta_1 \left(\frac{1}{x}\right)$



Try: E(ln(Y)) = $\beta_0 + \beta_1 x$



Try: E(Y) = $\beta_0 + \beta_1 ln(x)$



Try: E(ln(Y)) = $\beta_0 + \beta_1 ln(x)$

NORMALITY OF ERROR TERM (NOISE)

Lecture 4c-Part 2

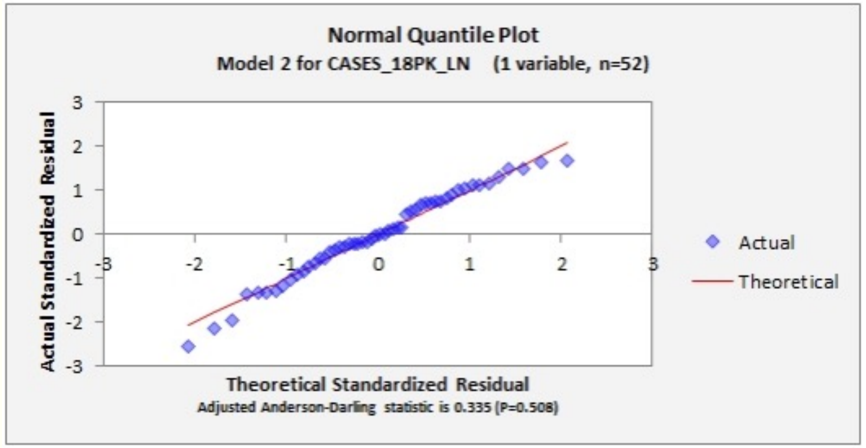
Violations of Assumptions – Normality of Noise

- In regression, the noise term is assumed to be normally distributed with a mean of zero.
- **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 $\boldsymbol{e} = (y - \hat{y})$
- Violations of normality create problems for determining whether:
  - model coefficients are significantly different from zero, and
  - for calculating confidence intervals for forecasts,
- If the error distribution is significantly non-normal
  - probabilities of Type I and Type II errors may be incorrect in hypothesis tests, leading to incorrect conclusions, and
  - confidence intervals may be too wide or too narrow
- *However*, if the only goal is to estimate its coefficients and *generate predictions* in such a way as to minimize mean squared error, then normality violations are typically not a threat.
- **Causes of non-Normality:** violations of normality often arise either because
  a) *the distributions of the dependent and/or independent variables are themselves significantly non-normal,*
  b) *the linearity assumption (between dependent and independent variable(s) is violated*, and/or
  c) there are a few unusual data points or "outliers" that "skew" the distribution and should be studied closely. Since parameter estimation is based on the minimization of squared error, a few extreme observations can exert a disproportionate influence on parameter estimates.

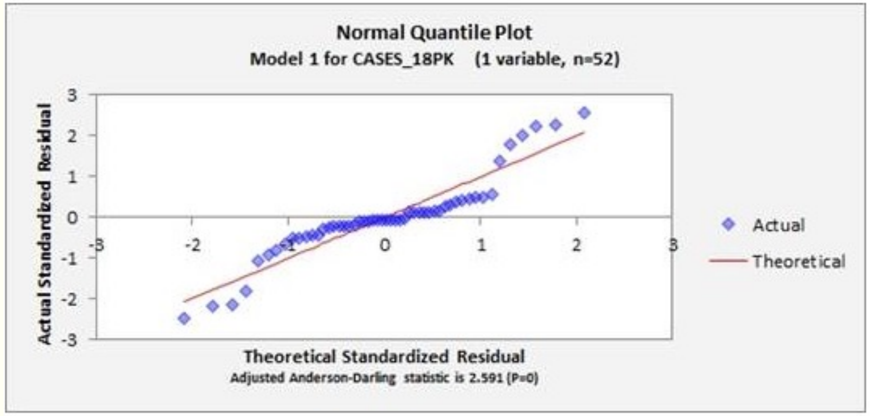
Violations of Assumptions – Normality of Residuals

- A visual check for normality of the noise term is through a *normal probability plot or normal quantile plot (also known as Q-Q plot)* of the residuals.
- These are plots of the fractiles or quantiles of error distribution versus the fractiles or quantiles of a normal distribution having the same mean and variance.
- If the distribution is normal, the points on such a plot should fall close to the diagonal reference line.
- A bow-shaped pattern of deviations from the diagonal indicates that the residuals have greater *skewness* ("asymmetry") than a normal distribution i.e., too many large values in one direction.
- An S-shaped pattern of deviations indicates that the variable has greater *kurtosis* ("heaviness of tails") than the normal distribution -- i.e., there are either too many or too few large values in both directions (positive and negative).
- Sometimes the problem is revealed to be that there are a few data points on one or both ends that deviate significantly from the reference line ("outliers"), in which case they should get close attention.

Residuals Approximately Normal



Residuals Deviating from Normal



Violations of Assumptions – Normality of Residuals

- However, we are mostly interested in the normality of the **residuals**. It is standard practice to **standardize** the residuals before analyzing them.
  - Remember that the residuals are expected to have a mean of 0. So in standardizing residuals, we simply divide them by their standard deviation (i.e., √MSE (Mean Squared Error)

- For the two models, we obtain the Q-Q Plot, Histogram and Skewness and Kurtosis for the standardized residuals from mod1. Skewness and Kurtosis are obtained using **library(moments)** we have used earlier.

- In R, the function *qqnorm()* gives the normal quantile plot and *qqline()* overlays the normal quantiles line on the plot.

mod1: $\widehat{\text{assets}} = 69.94 + 0.273\text{age} + 0.323\text{mortgage}$

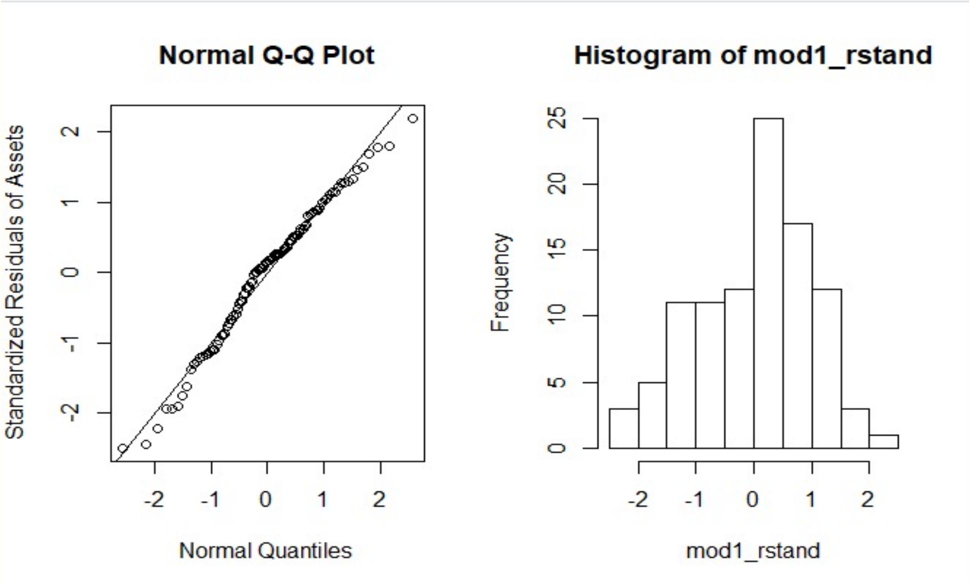
mod2: $\widehat{\text{assets}} = 44.421 + 0.192\text{age} + 13.695\text{logMort} + 0.804\text{logHome}$

Standardized Residuals

```
#
# Comparison of Standardized Residuals from Untransformed and Transformed Predictors
#
library(moments)
mod1_rstand <-rstandard(mod1)
qqnorm(mod1_rstand, ylab="Standardized Residuals of Assets", xlab="Normal Quantiles")
qqline(mod1_rstand)
hist(mod1_rstand)
print(skewness(mod1_rstand))
print(kurtosis(mod1_rstand))
#
mod2_rstand <-rstandard(mod2)
qqnorm(mod2_rstand, ylab="Standardized Residuals of Assets", xlab="Normal Quantiles")
qqline(mod2_rstand)
hist(mod2_rstand)
print(skewness(mod2_rstand))
print(kurtosis(mod2_rstand))
#
```

Violations of Assumptions – Normality of Residuals

- For mod1, the histogram shows that the standardized residuals are somewhat skewed to the left (the skewness is negative at -0.375) and it is somewhat platykurtic (Kurtosis is less than 3 at 2.696)
- The Q-Q plot is somewhat bow-shaped (indicating some skewness) and also S-shaped (crossing the line in both directions) indicating Kurtosis different from a normal distribution.




```
> print(skewness(mod1_rstand))
[1] -0.3751965
> print(kurtosis(mod1_rstand))
[1] 2.696205
```

mod1: $\widehat{\text{assets}} = 69.94 + 0.273\text{age} + 0.323\text{mortgage}$
mod2: $\widehat{\text{assets}} = 44.421 + 0.192\text{age} + 13.695\text{logMort} + 0.804\text{logHome}$

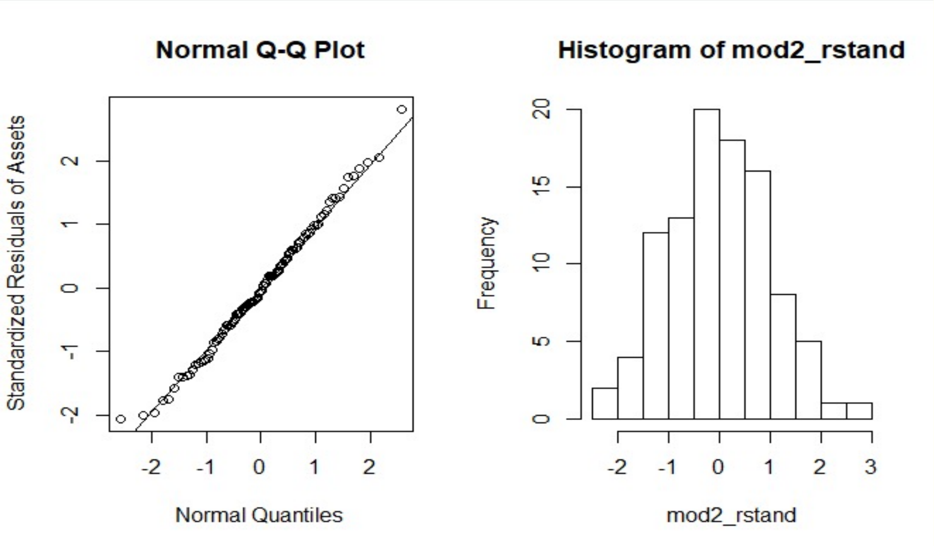
Violations of Assumptions – Normality of Residuals

- For mod2, the histogram shows that the standardized residuals are somewhat skewed to the right (the skewness is positive at -0.169) and it is somewhat platykurtic (Kurtosis is less than 3 at 2.719)

- The Q-Q plot is less bow-shaped than for mod1 (indicating some skewness) and is also S-shaped (crossing the line in both directions) indicating Kurtosis different from a normal distribution.

- Overall, the standardized residuals for mod2 seem closer to normality than those from mod1.

- Neither model appears to show drastic deviations of residuals from normality.




```
> print(skewness(mod2_rstand))
[1] 0.1686719
> print(kurtosis(mod2_rstand))
[1] 2.718532
```

mod1: $\widehat{assets} = 69.94 + 0.273age + 0.323mortgage$
mod2: $\widehat{assets} = 44.421 + 0.192age + 13.695logMort + 0.804logHome$

Violations of Assumptions – Tests for Normality of Residuals

- There are also a variety of **statistical tests** for normality, including the Kolmogorov-Smirnov test, the Shapiro-Wilk test, the Jarque-Bera test, and the Anderson-Darling test.
- Some of these test results change depending on whether raw residuals or standardized residuals are tested.
- The Anderson-Darling test is generally considered to be the best, because it is specific to the normal distribution (unlike the K-S test) and it looks at the whole distribution rather than just the skewness and kurtosis (like the J-B test).
- However, many of these tests are sensitive to sample sizes, failing to detect deviations from normality for small samples, and more easily **rejecting** the null hypothesis of normality for large samples.
- It is usually better to focus more on violations of the other assumptions and/or the influence of a few outliers (which may be mainly responsible for violations of normality anyway) and to look at a normal probability plot or normal quantile plot and draw your own conclusions about whether the problem is serious and whether it is systematic.

Violations of Assumptions – Tests for Normality of Residuals

- The null hypothesis of these tests is that "sample distribution is normal". If the test is **significant**, the distribution is non-normal.
- In our case, neither test rejects the null hypothesis that the standardized residuals are normal at $\alpha = 0.05$.
- Our sample size is reasonable (100) so in our case we can assume that the residuals are normal in both models.

```
> # Kolomogorov-Smirnov Test
> print(ks.test(mod1_rstand,"pnorm"))

        One-sample Kolmogorov-Smirnov test

data:  mod1_rstand
D = 0.08968, p-value = 0.3972
alternative hypothesis: two-sided

> print(ks.test(mod2_rstand, "pnorm"))

        One-sample Kolmogorov-Smirnov test

data:  mod2_rstand
D = 0.043257, p-value = 0.9921
alternative hypothesis: two-sided

> # Shapiro-Wilk Test
> print(shapiro.test(mod1_rstand))

        Shapiro-Wilk normality test

data:  mod1_rstand
W = 0.98146, p-value = 0.1725

> print(shapiro.test(mod2_rstand))

        Shapiro-Wilk normality test

data:  mod2_rstand
W = 0.99274, p-value = 0.8717
```

mod1: $\widehat{\text{assets}} = 69.94 + 0.273\text{age} + 0.323\text{mortgage}$

mod2: $\widehat{\text{assets}} = 44.421 + 0.192\text{age} + 13.695\text{logMort} + 0.804\text{logHome}$

HOMOSCEDASTICITY OF ERROR TERM (NOISE)

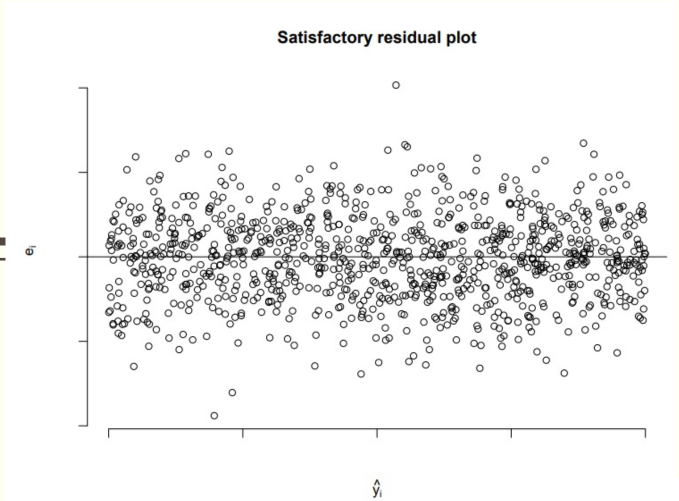
Lecture 4c-Part 3

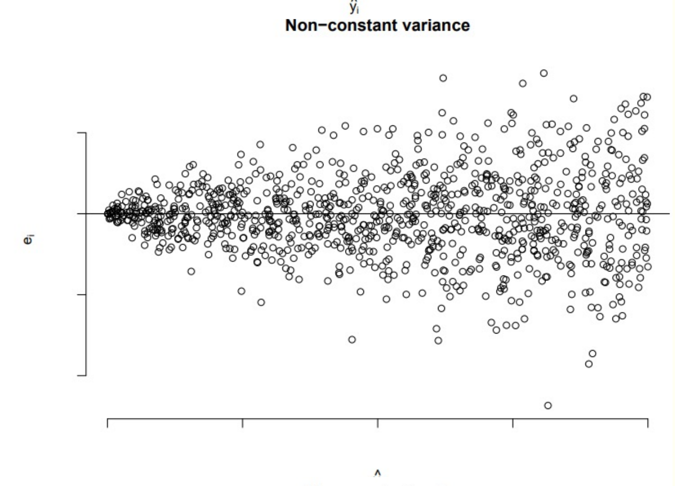
Violations of Assumptions – Homoscedasticity/Heteroscedasticity

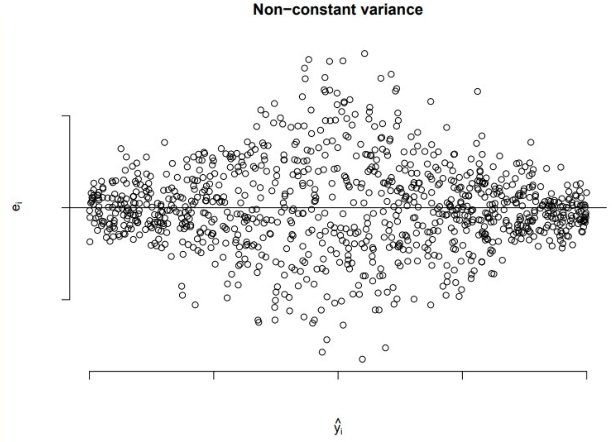
- In regression, the noise term is assumed to have constant variance that is independent of X (i.e., it is homoscedastic)
- Population Model:
  - $Y = \boldsymbol{\alpha} + \boldsymbol{\beta_1} X_1 + \boldsymbol{\beta_2} X_2 + \boldsymbol{\beta_3} X_3 + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\sigma_\varepsilon}^2)$.
- (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 $\boldsymbol{e} = (y - \hat{y})$
- A violation of homoscedasticity ("heteroscedasticity") may have the effect of giving too much weight (or too little weight) to a small subset of the data (namely the subset where the error variance was largest or smallest) when estimating coefficients.
- Heteroscedasticity makes it difficult to gauge the true standard deviation of the prediction errors, usually resulting in confidence intervals (for expected value of Y) that are too wide or too narrow.
- Heteroscedasticity can be a byproduct of a significant violation of the linearity and/or independence assumptions, in which case it may also be fixed as a byproduct of fixing those problems.

Violations of Assumptions – Homoscedasticity/Heteroscedasticity

- **How to diagnose**:
  - In a plot of **<u>residuals versus predicted values</u>** look for residuals that grow larger either as a function of the independent variable or predicted value.
  - Because of imprecision in the coefficient estimates, the errors may tend to be slightly larger for forecasts associated with predictions or values of independent variables that are extreme in both directions, although the effect should not be too dramatic. What you hope not to see are errors that systematically get larger in one direction by a significant amount.
- **How to fix**:
  - If the dependent variable is strictly positive and if the residual-versus-predicted plot shows that the size of the errors is proportional to the size of the predictions (i.e., if the errors seem consistent in percentage rather than absolute terms), a *log transformation* applied to the dependent variable may be appropriate.





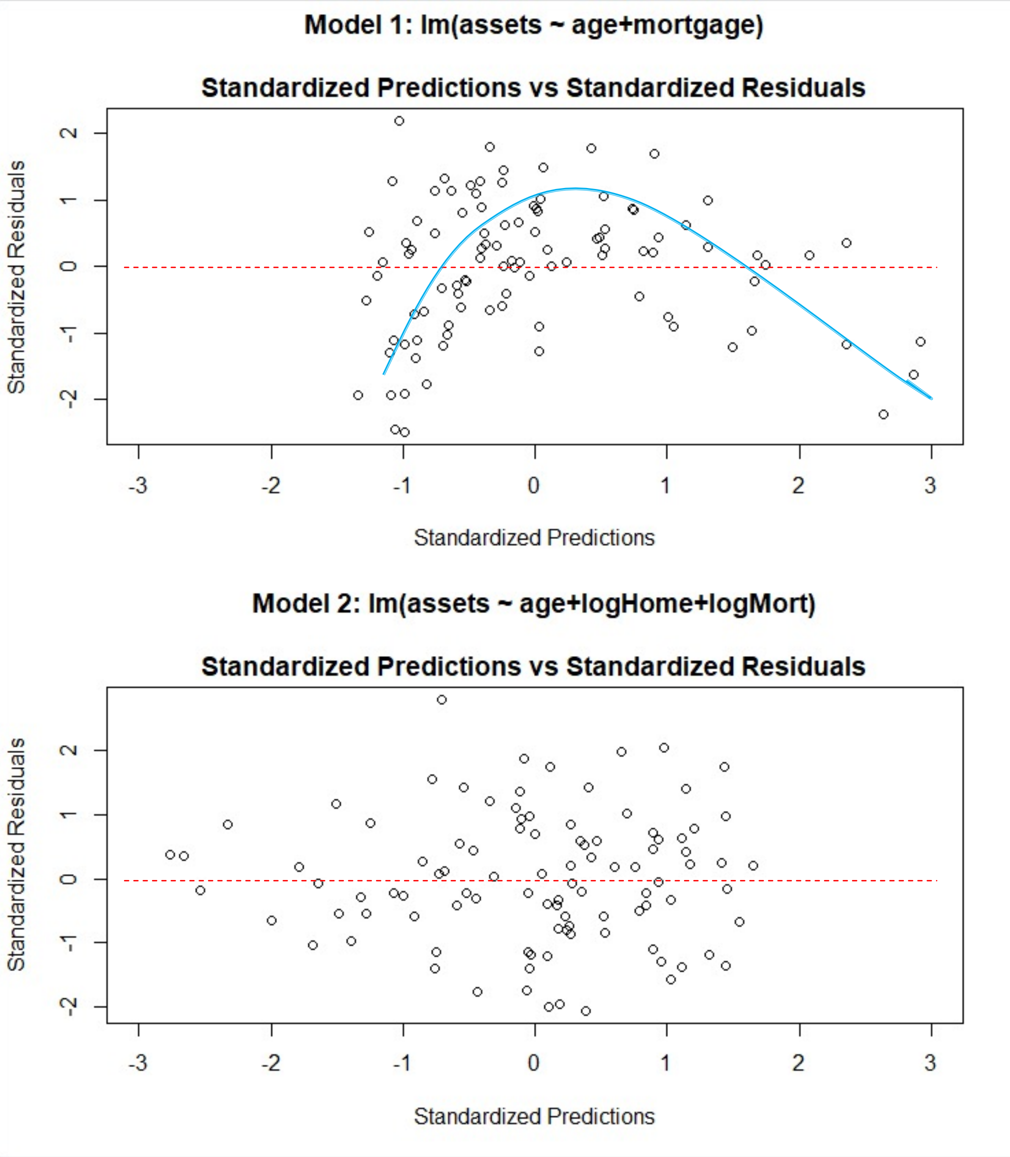


Violations of Assumptions – Homoscedasticity/Heteroscedasticity

mod1: $\widehat{\text{assets}} = 69.94 + 0.273\text{age} + 0.323\text{mortgage}$
mod2: $\widehat{\text{assets}} = 44.421 + 0.192\text{age} + 13.695\text{logMort} + 0.804\text{logHome}$

- We show the plots of Standardized Residuals vs. Standardized Predicted values for each of the two models.
- We can see clear heteroscedasticity for the plot for mod1 where the spread is uneven around the horizontal zero line, for all standardized predictions on the x-axis.
- The plot for mod2 shows relatively even spread (accounting for fewer observations at the lower end) around the zero line.

```
> #
> # Plots for Checking Heteroscedascticity
> # using Standardized Residuals vs Standardized Predictions
> #
> # Getting Standardized Predictions for Each Model
> #
> mod1_pred <- predict(mod1)
> mod2_pred <- predict(mod2)
> mod1_pstand <-  (mod1_pred - mean(mod1_pred))/sd(mod1_pred)
> mod2_pstand <-  (mod2_pred - mean(mod2_pred))/sd(mod2_pred)
> #
> # Do the plot for each model
> #
> par(mfrow=c(2,1))
> plot(mod1_pstand, mod1_rstand,
+      main = "Model 1: lm(assets ~ age+mortgage)
+      \nStandardized Predictions vs Standardized Residuals",
+      xlab = "Standardized Predictions",
+      ylab = "Standardized Residuals",
+      xlim = c(-3, 3))
>
> plot(mod2_pstand, mod2_rstand,
+      main = "Model 2: lm(assets ~ age+logHome+logMort)
+      \nStandardized Predictions vs Standardized Residuals",
+      xlab = "Standardized Predictions",
+      ylab = "Standardized Residuals",
+      xlim = c(-3, 3))
> |
```

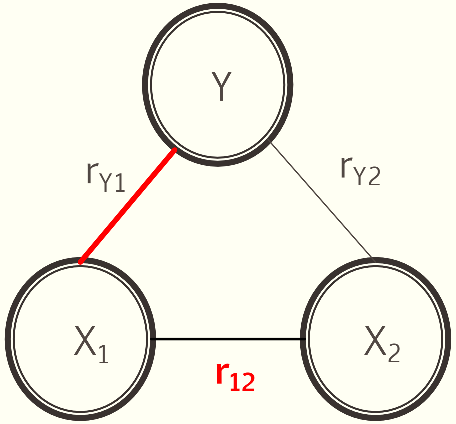


MULTICOLLINEARITY

Lecture 4c-Part 4

Multicollinearity

- Multicollinearity refers to the situation where two or more of the independent (X) variables (predictors or regressors) are highly correlated with each other.
  - $Y = \boldsymbol{\alpha} + \boldsymbol{\beta_1} X_1 + \boldsymbol{\beta_2} X_2 + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\sigma_\varepsilon}^2)$.
  - $\hat{\beta}_1 = \frac{r_{y1} - r_{y2} * r_{12}}{1 - r_{12}^2}$ and $\hat{\beta}_2 = \frac{r_{y2} - r_{y1} * r_{12}}{1 - r_{12}^2}$



- That is, multicollinearity *means that* r_{12} *is very high.* You can see that as r_{12} gets closer to 1, the estimates of the slope ($\hat{\beta}$) coefficients get **inflated** and **unstable**.
- **Inflation** means that when you introduce X_1 and X_2 together in the same model, both the $\hat{\beta}$ estimates rise dramatically. This is often a sure sign of multicollinearity.
  - Intuitively, If X_1 and X_2 are highly correlated, they are the "same variable"; controlling for one of them (say X_2), produces very little variation in the other variable (X_1). Since "slope" is just "change in dependent (Y) for unit change in independent (X_1)", when there is very little variation in the X_1 variable controlling for the correlation between X_1 and X_2, the slope $\hat{\beta}_1$ increases dramatically. This will also be true for $\hat{\beta}_2$.
- **Unstable** means
  - the $\hat{\beta}_1$(when used without X_2 in the model) changes dramatically when X_2 is included (or vice versa). (OR)
  - $\hat{\beta}_2$remains positive and $\hat{\beta}_1$flips sign (Note: Not all "sign flips" are due to Multicollinearity)
- Multicollinearity also inflates the **standard error** of the slope of X_1 so that the t-statistic ($\hat{\beta}_1$ /standard error) of ($\hat{\beta}_1$)) becomes "**unstable**" and not statistically significant.

Multicollinearity -Demonstration

```
# Multicollinearity demonstration
#
# # From Rosenkrantz, "Probability and Statistics for Science, Engineering
# and Finance," CRC Press, Boca raton, 2009. Table 10.2.
# 12 1992 cars were measured for fuel efficiency. The response variable
# is miles per gallon (MPG).
```

```
> car_weight <- c(2.495, 2.53, 2.62, 3.395, 3.03, 3.345, 3.04, 3.085, 3.495, 3.95, 3.47, 4.105) # weight in 1000 pounds
> car_mpg <- c(32, 30, 29, 25, 27, 28, 29, 27, 28, 25, 28, 25) # Miles per gallon
> car_disp <- c(1.9, 1.8, 1.6, 3, 2.2, 3.8, 2.2, 3, 3.8, 4.6, 3.8, 5.7) # Engine displacement in liters
> #
> M <- cbind(car_mpg, car_weight, car_disp)
> print(cor(M))
              car_mpg car_weight   car_disp
car_mpg     1.0000000 -0.8318262 -0.6997975
car_weight -0.8318262  1.0000000  0.9534827
car_disp   -0.6997975  0.9534827  1.0000000
> #
> print("car_weight and car_disp are highly correlated - high collinearity")
[1] "car_weight and car_disp are highly correlated - high collinearity"
> #
```

As car weight and engine displacement increase, car_mpg decereases

Multicollinearity -Demonstration

- When we use car_disp alone to predict car_mpg (Model car_mod1) $\hat{\beta}_{disp}$ is -1.1859 and significant at α = 0.05 with p-value of 0.0113
- When we introduce car_weight (which has high correlation with car_disp) as a predictor into the model car_mod2:
    - The sign of $\hat{\beta}_{disp}$ becomes positive (which is incorrect)
    - $\hat{\beta}_{disp}$ is no longer a significant predictor at α = 0.05 with p-value of 0.07445.
    - Also note that the standard error of $\hat{\beta}_{disp}$ increased from 0.3828 in mod1 to 0.8632 in mod2.
- *<u>What is the recommended fix</u>?* You should not use both car_weight and car_disp as predictors in the model car_mod2 because they are almost the "same variable". Use one or the other. (There are other options, but they are beyond the scope of this course)

```
> car_mod1 <- lm(car_mpg ~ car_disp)
> car_mod2 <- lm(car_mpg ~ car_weight+car_disp)
> #
> summary(car_mod1)

Call:
lm(formula = car_mpg ~ car_disp)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8884 -0.9140  0.2383  1.0604  2.8071 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  31.4462     1.2795  24.576 2.84e-10 ***
car_disp     -1.1859     0.3828  -3.098   0.0113 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.601 on 10 degrees of freedom
Multiple R-squared:  0.4897,	Adjusted R-squared:  0.4387 
F-statistic: 9.597 on 1 and 10 DF,  p-value: 0.01129

> summary(car_mod2)

Call:
lm(formula = car_mpg ~ car_weight + car_disp)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.5065 -0.5705 -0.2401  0.9839  1.5496 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  46.3513     4.2812  10.827 1.84e-06 ***
car_weight   -7.4770     2.1029  -3.556  0.00616 ** 
car_disp      1.7406     0.8632   2.016  0.07455 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.088 on 9 degrees of freedom
Multiple R-squared:  0.7878,	Adjusted R-squared:  0.7406 
F-statistic: 16.71 on 2 and 9 DF,  p-value: 0.000934
```

Multicollinearity - Detection

- Some signs of Multicollinearity include:
  - A regression coefficient is not significant even though, theoretically, that variable should be highly correlated with Y.
  - When you add or delete an X variable, the regression coefficients change dramatically.
  - You see a negative regression coefficient when your response should *increase* along with X.
  - You see a positive regression coefficient when the response should *decrease* as X increases.
  - X variables have high pairwise correlations (such as say > 0.8). However, this can be misleading because Multicollinearity may involve more than the pairwise relationship between two variables, such as the effect of a third variable on X_1 and X_2.

Multicollinearity – Detection

- It is better to rely on **Multicollinearity diagnostics**.
- **Tolerance**: If you have 3 independent variables: X1, X2, X3, then Tolerance is based on doing a regression: X1 is dependent; X2 and X3 are independent and computing Tolerance for X1 as (1 - regression R^2). A tolerance value lower than 0.2 typically indicates potential multicollinearity.
- Alternatively, **VIF** (**variance Inflation Factor**) which is 1/Tolerance can also be used.
  - If the VIF is equal to 1 there is no multicollinearity, but a VIF between 5 and 10 indicates high correlation of X_1 with one or more of the other predictors, and may indicate potential Multicollinearity problems when X_1 is included with those predictors.
- In **R**, we use library (olsrr) to get Tolerance and VIF. We note that Tolerance is less than 0.2 and VIF is greater than 5, incidcating multicollinearity between car_weigt and car_disp.

```
> #
> # Collinearity diagnostics using library "olsrr"
> library(olsrr)
> ols_vif_tol(car_mod2)
# A tibble: 2 x 3
  Variables  Tolerance   VIF
  <chr>          <dbl> <dbl>
1 car_weight    0.0909  11.0
2 car_disp      0.0909  11.0
```

Multicollinearity

- For our models:

  mod1: $\widehat{\text{assets}} = 69.94 + 0.273\text{age} + 0.323\text{mortgage}$
  mod2: $\widehat{\text{assets}} = 44.421 + 0.192\text{age} + 13.695\text{logMort} + 0.804\text{logHome}$

- There are no issues with multicollinearity

```
> #
> ols_vif_tol(mod1)
# A tibble: 2 x 3
  Variables Tolerance   VIF
  <chr>         <dbl> <dbl>
1 age           0.841  1.19
2 mortgage      0.841  1.19
> ols_vif_tol(mod2)
# A tibble: 3 x 3
  Variables Tolerance   VIF
  <chr>         <dbl> <dbl>
1 age           0.747  1.34
2 logHome       0.431  2.32
3 logMort       0.496  2.01
```

Multicollinearity

- If the regression model is used *primarily for prediction*, like violation of the normality of noise, multicollinearity does not create a problem
- Multicollinearity and violation of normality of noise affect the values of betas and their tests of significance and therefore become problematic when the model is used for deciding which predictors are important to the model.

OUTLIER ANALYSIS

Lecture 4c-Part 5

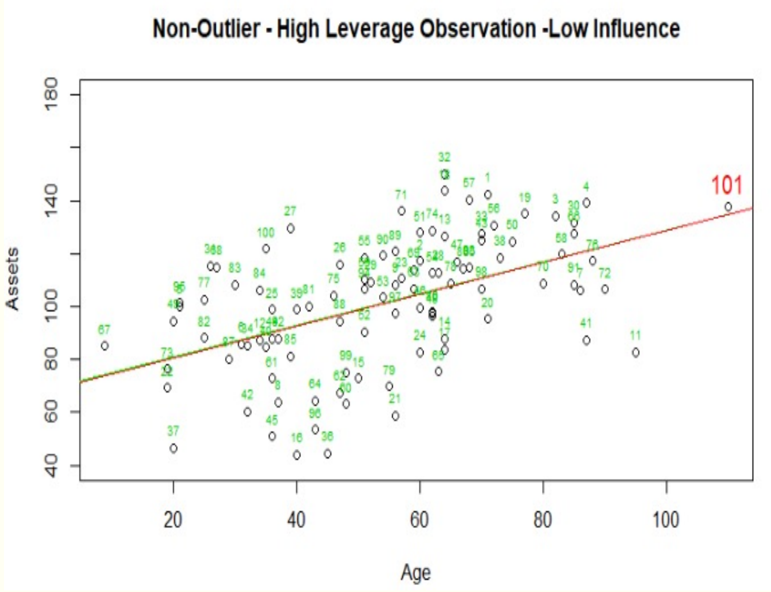
Outliers, High Leverage and Influential Observations

- An **outlier** is a data point whose response *y* does not follow the general trend of the rest of the data.
- A data point has high **leverage** if it has "extreme" predictor *x* values. With a single predictor, an extreme *x* value is simply one that is particularly high or low. With multiple predictors, extreme x values may be particularly high or low for one or more predictors, or may be "unusual" combinations of predictor values (e.g., with two predictors that are positively correlated, an unusual combination of predictor values might be a high value of one predictor paired with a low value of the other predictor).
  - For our purposes, we consider a data point to be an outlier *only if* it is extreme with respect to the other *y* values, not the *x* values.
- A data point is **influential** if it unduly influences any part of a regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results. Outliers and high leverage data points have the *potential* to be influential, but we generally have to investigate further to determine whether or not they are actually influential.

https://newonlinecourses.science.psu.edu/stat462/node/170/

Outliers, High Leverage and Influential Observations

- I have added a point with ID = 101 with different values of age vs assets to illustrate outliers.
- In this first example, obs. 101 has *high leverage* but is *not an outlier*. This is because obs. 101 is in line with the trend of the other observations and has minimal influence.
- You can see from the regressions with and without obs. 101 that the regression outputs are unchanged. Its *influence is low*.
- R^2, the intercept, and the estimated slope of age (or its significance) are not changed much.



Without obs. 101

```
Call:
lm(formula = assets ~ age, data = df_age)

Residuals:
    Min      1Q  Median      3Q     Max 
-51.297 -11.122   4.684  16.397  42.991 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  68.7658     6.2337  11.031  < 2e-16 ***
age           0.5996     0.1093   5.484 3.25e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.21 on 98 degrees of freedom
Multiple R-squared:  0.2348,	Adjusted R-squared:  0.227 
F-statistic: 30.07 on 1 and 98 DF,  p-value: 3.249e-07
```

With obs. 101

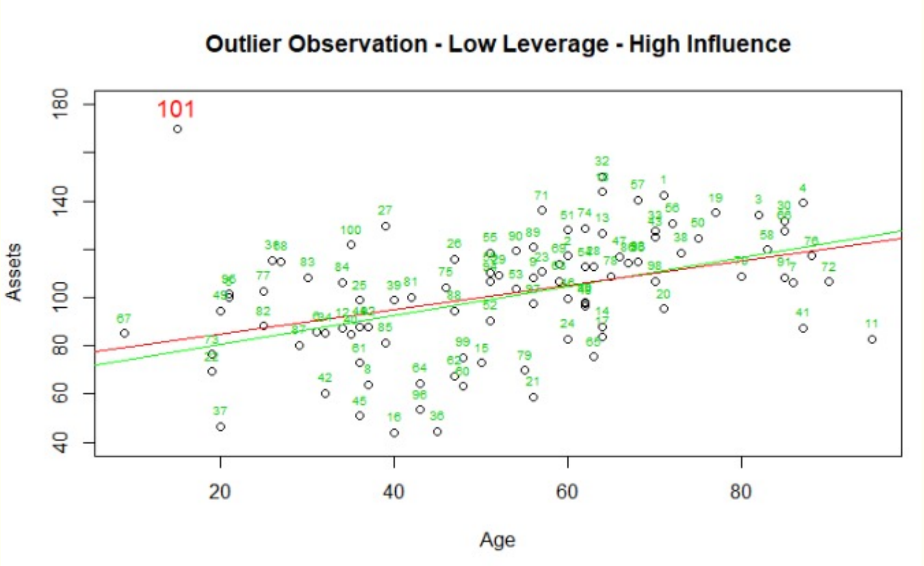
```
Call:
lm(formula = assets ~ age, data = df_age1)

Residuals:
    Min      1Q  Median      3Q     Max 
-51.288 -10.974   4.453  16.061  42.914 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  68.5551     6.0384   11.35  < 2e-16 ***
age           0.6041     0.1045    5.78  8.7e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.11 on 99 degrees of freedom
Multiple R-squared:  0.2523,	Adjusted R-squared:  0.2447 
F-statistic: 33.41 on 1 and 99 DF,  p-value: 8.703e-08
```

Outliers, High Leverage and Influential Observations

- In this second example, obs. 101 has *low leverage* and appears out of sync with the general trend of the data. It is a *potential outlier.*

- You can see from the regressions with and without obs. 101 that the regression outputs have changed. Hence, *this potential outlier is influential.*

- R^2, the intercept, and the estimated slope of age are different, and we may get different predictions.




```
Call:
lm(formula = assets ~ age, data = df_age)

Residuals:
    Min      1Q  Median      3Q     Max 
-51.297 -11.122   4.684  16.397  42.991 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  68.7658     6.2337  11.031  < 2e-16 ***
age           0.5996     0.1093   5.484 3.25e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.21 on 98 degrees of freedom
Multiple R-squared:  0.2348,	Adjusted R-squared:  0.227 
F-statistic: 30.07 on 1 and 98 DF,  p-value: 3.249e-07
```

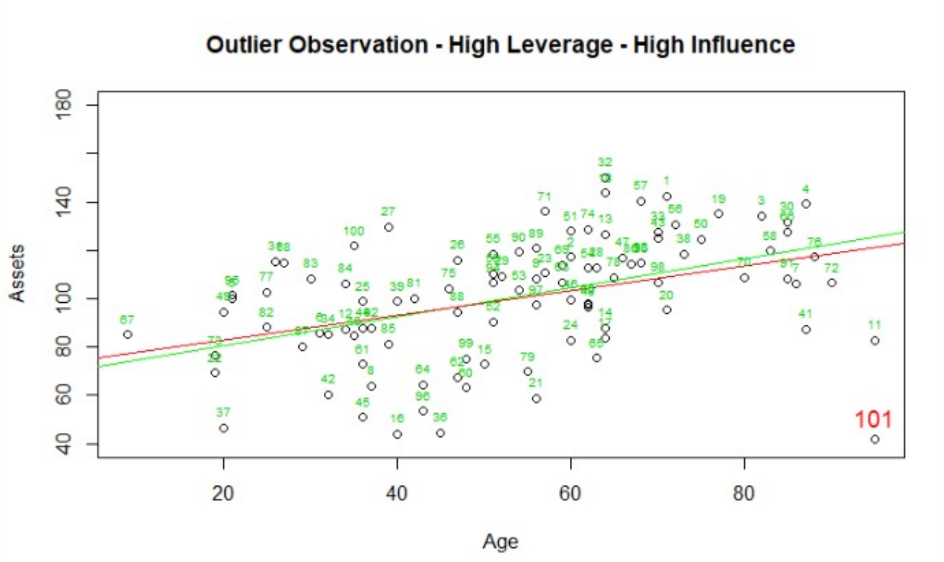
Without obs. 101

```
Call:
lm(formula = assets ~ age, data = df_age2)

Residuals:
    Min      1Q  Median      3Q     Max 
-52.952 -12.089   4.055  15.062  87.881 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  74.4774     6.5890  11.303  < 2e-16 ***
age           0.5094     0.1161   4.388 2.87e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.96 on 99 degrees of freedom
Multiple R-squared:  0.1628,	Adjusted R-squared:  0.1543 
F-statistic: 19.25 on 1 and 99 DF,  p-value: 2.867e-05
```

With obs. 101

Outliers, High Leverage and Influential Observations

- In this third example, obs. 101 has *high leverage* and appears out of sync with the general trend of the data. It is a *potential outlier.*
- You can see from the regressions with and without obs. 101 that the regression outputs have changed. Hence, *this potential outlier is influential.*
- R^2, the intercept, and the estimated slope of age are different, and we may get different predictions.



Without obs. 101

```
Call:
lm(formula = assets ~ age, data = df_age)

Residuals:
    Min      1Q  Median      3Q     Max 
-51.297 -11.122   4.684  16.397  42.991 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  68.7658     6.2337  11.031  < 2e-16 ***
age           0.5996     0.1093   5.484 3.25e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.21 on 98 degrees of freedom
Multiple R-squared:  0.2348,	Adjusted R-squared:  0.227 
F-statistic: 30.07 on 1 and 98 DF,  p-value: 3.249e-07
```

With obs. 101

```
Call:
lm(formula = assets ~ age, data = df_age2)

Residuals:
    Min      1Q  Median      3Q     Max 
-79.322 -11.491   4.646  16.602  44.691 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  72.6487     6.5648  11.066  < 2e-16 ***
age           0.5124     0.1141   4.489 1.94e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.64 on 99 degrees of freedom
Multiple R-squared:  0.1691,	Adjusted R-squared:  0.1607 
F-statistic: 20.15 on 1 and 99 DF,  p-value: 1.94e-05
```

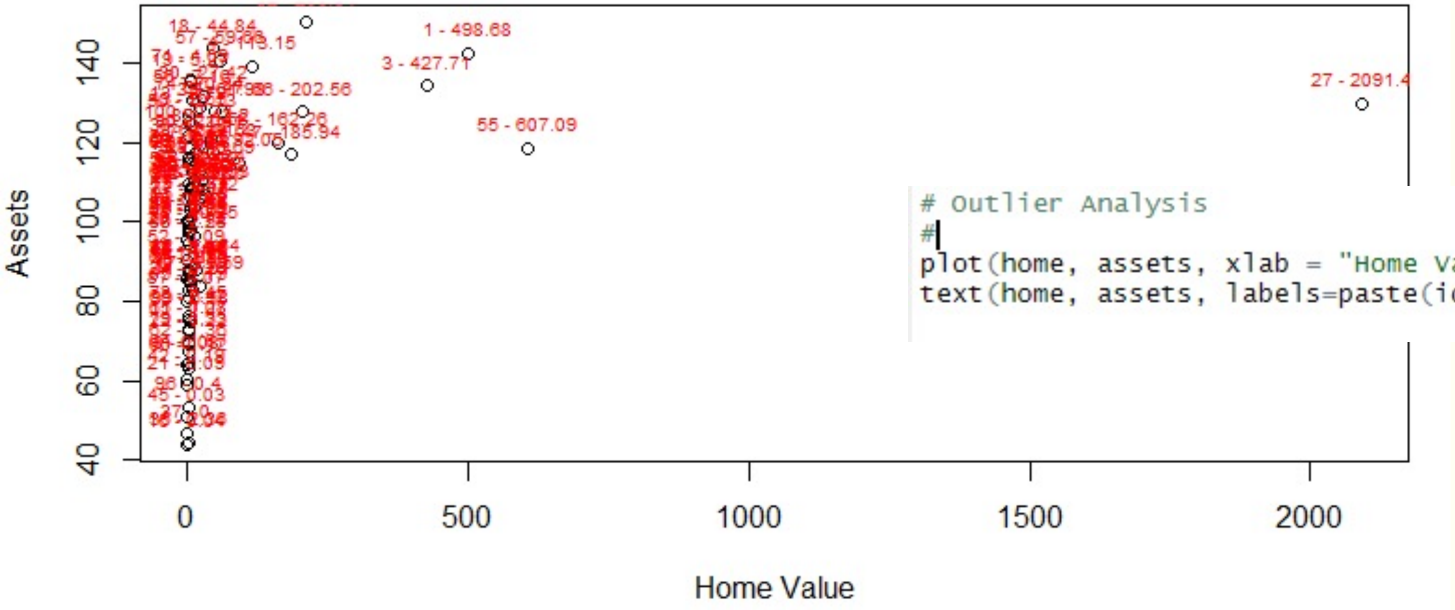
Outlier Analysis of Observations

- In practice, when we have many predictors, identifying influential outliers involves a multi-dimensional analysis. Visual identification from bi-variate plots with one predictor at a time can be misleading.

- Instead, we rely on *outlier diagnostics*, where the influence of every observation is assessed taking into account all predictors in the model.

- Earlier, we saw (prior to the log transformation) that a plot of Assets vs. Home value showed one value of Home which appears to be “extreme”. This was the observation with obs = 27 and had a value of 2091.4.

- The question becomes, “Is this an Influential Outlier?” That is, does its inclusion and exclusion




```
# Outlier Analysis
#
plot(home, assets, xlab = "Home Value", ylab = "Assets")
text(home, assets, labels=paste(id, "-", home), cex = 0.6, pos = 3, col = 2)
```

Outlier Diagnostic Measures

- The **Leverage** of an observation refers to an *extreme value on a predictor variable* (“*X axis value*”)
  - Leverage is a measure of how far an independent variable deviates from its mean.
  - These leverage points can be influential i.e., have an effect on the estimate of regression coefficients.
- An **Outlier** is an observation with large residual.
  - An observation whose *residual value (“Y axis value”)* is unusual given its values on the predictor variables.
  - An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.
- The **Influence** of an observation can be thought of as a combination of leverage and outlierness.
  - Removing the observation substantially changes the estimate of coefficients.
- The corresponding measures are:
  - r_i = Studentized residual - - for measuring “outlierness” or extremeness on the Y-axis
  - h_i = Leverage - for measuring “unusualness” of each x-value, relative to the all other x-values
  - D_i = Cook’s distance - for measuring influence

Outlier Diagnostic Measures - Leverage

- **Leverage** (**h** or **hat value**):
  - It measures the distance between the x_i for i^{th} case (observation), and the average of X
    - $h_i = \frac{1}{(n-1)}\left(\frac{(x_i - \bar{x})^2}{s_x}\right) + \frac{1}{n}$
    - i.e., it is the proportion of the total sum of squares of X that is attributed to the observation x_i
    - In the plot shown, observation ID=27 is the most extreme *along the x-axis* and has the most leverage

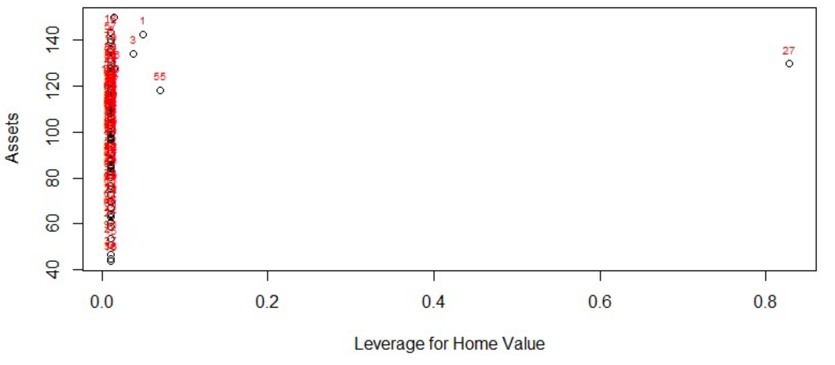
  - The further away from the mean of X (either in a positive or negative direction), the more leverage an observation has on the regression fit

```
> mod_home <- lm(assets~home)
> summary(mod_home)

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

> leverage <- hatvalues(mod_home)
> stud_res <- rstudent(mod_home)
> cook_dist <- cooks.distance(mod_home)
> 
> #
> df_home <-data.frame(home, assets, leverage, stud_res, cook_dist)
> print(round(df_home,4))
```

Outlier Diagnostic Measures - Leverage

- **Leverage** (**h** or **hat value**):
  - The value of the leverage for observation 27 was 0.8281. This is clearly very high relative to all other Home-values. If every observation has the same leverage, its value will be 1/n or 0.01. So relative to all other values, observation 27 has very high leverage
  - High leverage does not necessarily mean that it influences the regression coefficients
    - It is possible to have a high leverage and yet follow straight in line with the pattern of the rest of the data
  - High leverage also *does not mean* the observation is an

```
> #
> plot(leverage, assets, xlab = "Leverage for Home Value", ylab = "Assets")
> text(leverage, assets, labels=paste(id), cex = 0.6, pos = 3, col = 2)
> |
```




```
> #
> df_home <-data.frame(home, assets, leverage, stud_res, cook_dist)
> print(round(df_home,4))
```

	home	assets	leverage	stud_res	cook_dist
1	498.68	142.49	0.0490	1.2692	0.0412
2	16.21	117.29	0.0103	0.7502	0.0029
3	427.71	134.15	0.0375	0.9825	0.0188
4	113.15	139.08	0.0107	1.5823	0.0133
5	0.07	100.01	0.0106	0.0279	0.0000
6	0.17	85.72	0.0106	-0.5853	0.0018
7	11.58	106.20	0.0104	0.2789	0.0004
8	0.08	64.02	0.0106	-1.5317	0.0124
9	25.20	108.16	0.0102	0.3462	0.0006
10	4.92	97.62	0.0105	-0.0804	0.0000
11	7.01	82.68	0.0104	-0.7247	0.0028
12	1.04	87.33	0.0106	-0.5171	0.0014
13	2.79	126.67	0.0105	1.1746	0.0073
14	3.57	87.73	0.0105	-0.5030	0.0014
15	1.07	72.81	0.0106	-1.1459	0.0070
16	0.04	43.93	0.0106	-2.4460	0.0304
17	21.59	83.81	0.0102	-0.6940	0.0025
18	44.84	143.63	0.0100	1.8735	0.0173
19	5.27	135.06	0.0105	1.5411	0.0124
20	0.19	95.27	0.0106	-0.1753	0.0002
21	0.09	58.77	0.0106	-1.7661	0.0163
22	1.22	69.31	0.0106	-1.2995	0.0089
23	24.84	110.59	0.0102	0.4509	0.0011
24	0.25	82.75	0.0106	-0.7134	0.0027
25	0.94	98.88	0.0106	-0.0216	0.0000
26	0.87	115.88	0.0106	0.7084	0.0027
27	2091.40	129.77	0.8281	-3.2019	22.5690
28	40.69	112.83	0.0100	0.5276	0.0014
29	0.20	109.37	0.0106	0.4290	0.0010

Outlier Diagnostic Measures – Studentized Residuals

- **Studentized residual (t_{res})** for detecting outliers:
  - One problem when trying to identify outliers, that the potential outlier may influence the regression model by "pulling" the estimated regression function towards the potential outlier, so that it isn't flagged as an outlier using the standardized residual criterion.
  - To address this issue, **studentized residuals** offer an alternative criterion for identifying outliers. The basic idea is to delete the observations one at a time, each time refitting the regression model on the remaining *n*–1 observations.
  - Then, we compare the observed response values to their fitted values based on the models with the i^{th} observation deleted.
  - This produces **deleted residuals**. Standardizing the deleted residuals produces **studentized residuals**.
  - $\mathbf{t_{res}} = e_i/\text{Std Err}(e_i) = \frac{e_i}{\sqrt{MSE_i\,(1-h_i)}}$, where MSE_i is the Mean Square Error of the regression model with the i^{th} observation deleted, and h_i is the leverage of the i^{th} observation.
  - An extreme studentized residual indicates an unusual predicted value for Y.

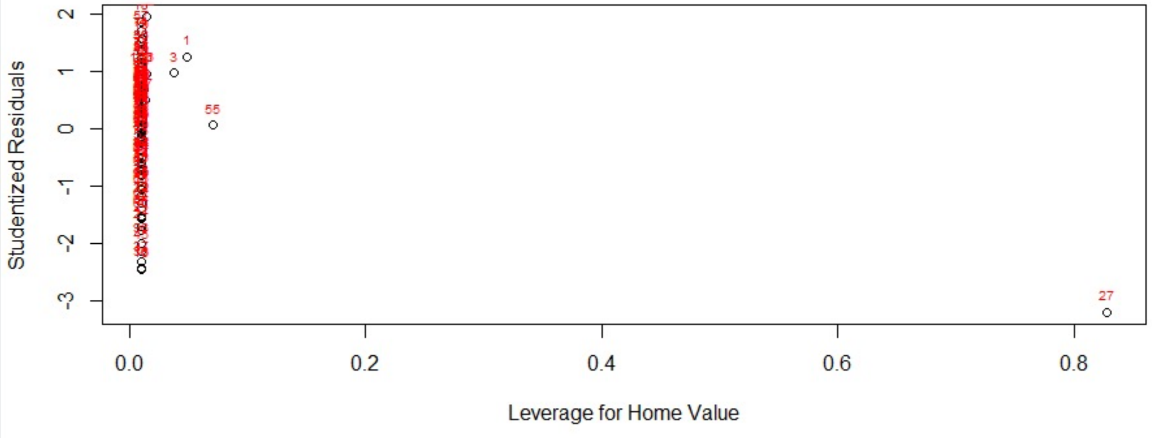
Outlier Diagnostic Measures – Studentized Residuals vs Leverage

- The studentized residual for observation 27 is **-3.06**, which is very high in magnitude relative to other observations and lies on the extreme tail of a t-distribution with n-k-2 degrees of freedom.

- This tells us that it has a very unusual predicted Y-value (we saw earlier it also had high leverage)

- The plot of the studentized residual vs leverage (hat-value) tells us:
  - Observation 27 has high leverage but has a low studentized residual
  - Observations 1 and 3 have relatively high studentized residual bout are low in leverage
  - Observation 55 has higher leverage but lower studentized residual than 1 and 3.

```
# Plot of Leverage (hat-value) versus Studentized Residuals
#
plot(leverage, stud_res, xlab = "Leverage for Home Value", ylab = "Studentized Residuals")
text(leverage, stud_res, labels=paste(id), cex = 0.6, pos = 3, col = 2)
```



	home	assets	leverage	stud_res	cook_dist
1	498.68	142.49	0.0490	1.2692	0.0412
2	16.21	117.29	0.0103	0.7502	0.0029
3	427.71	134.15	0.0375	0.9825	0.0188
4	113.15	139.08	0.0107	1.5823	0.0133
5	0.07	100.01	0.0106	0.0279	0.0000
6	0.17	85.72	0.0106	-0.5853	0.0018
7	11.58	106.20	0.0104	0.2789	0.0004
8	0.08	64.02	0.0106	-1.5317	0.0124
9	25.20	108.16	0.0102	0.3462	0.0006
10	4.92	97.62	0.0105	-0.0804	0.0000
11	7.01	82.68	0.0104	-0.7247	0.0028
12	1.04	87.33	0.0106	-0.5171	0.0014
13	2.79	126.67	0.0105	1.1746	0.0073
14	3.57	87.73	0.0105	-0.5030	0.0014
15	1.07	72.81	0.0106	-1.1459	0.0070
16	0.04	43.93	0.0106	-2.4460	0.0304
17	21.59	83.81	0.0102	-0.6940	0.0025
18	44.84	143.63	0.0100	1.8735	0.0173
19	5.27	135.06	0.0105	1.5411	0.0124
20	0.19	95.27	0.0106	-0.1753	0.0002
21	0.09	58.77	0.0106	-1.7661	0.0163
22	1.22	69.31	0.0106	-1.2995	0.0089
23	24.84	110.59	0.0102	0.4509	0.0011
24	0.25	82.75	0.0106	-0.7134	0.0027
25	0.94	98.88	0.0106	-0.0216	0.0000
26	0.87	115.88	0.0106	0.7084	0.0027
27	2091.40	129.77	0.8281	-3.2019	22.5690
28	40.69	112.83	0.0100	0.5276	0.0014
29	0.20	109.37	0.0106	0.4290	0.0010
30	27.42	131.68	0.0101	1.3632	0.0094
31	5.15	115.40	0.0105	0.6824	0.0025
32	209.91	150.13	0.0148	1.9581	0.0280
33	61.98	127.67	0.0100	1.1437	0.0066
34	3.53	85.36	0.0105	-0.6049	0.0020
35	9.58	114.55	0.0104	0.6402	0.0022
36	2.36	44.45	0.0105	-2.4247	0.0298
37	0.00	46.52	0.0106	-2.3251	0.0277
38	0.48	118.50	0.0106	0.8221	0.0036
39	1.65	98.91	0.0105	-0.0212	0.0000
40	4.66	84.44	0.0105	-0.6459	0.0022
41	1.98	87.24	0.0105	-0.5221	0.0015
42	0.19	60.45	0.0106	-1.6909	0.0150
43	2.55	125.20	0.0105	1.1107	0.0065
44	0.96	87.77	0.0106	-0.4980	0.0013
45	0.03	50.85	0.0106	-2.1253	0.0233
46	0.88	99.73	0.0106	0.0149	0.0000
47	185.94	116.79	0.0134	0.5195	0.0019
48	10.25	96.44	0.0104	-0.1375	0.0001
49	2.54	94.40	0.0105	-0.2154	0.0002
50	10.13	124.66	0.0104	1.0775	0.0061
51	47.00	127.85	0.0100	1.1704	0.0069
52	0.09	90.48	0.0106	-0.3806	0.0008
53	12.32	103.38	0.0103	0.1572	0.0001
54	8.22	112.67	0.0104	0.5609	0.0017
55	607.09	118.41	0.0703	0.0717	0.0002

Outlier Diagnostic Measures – Cook's D(istance)

- **Cook's Distance (D)**
- While hat values (leverage) measure the influence of X values and studentized residuals that of predicted Y values using residuals, both the *x* value and the *y* value of the data point play a role in the calculation of Cook's D.
- Like the studentized residual it measures the effect of deleting the i^{th} observation. However, it actually compares the slopes of the regression the two regression lines (with and without the i^{th} observation) and hence accounts for both x and y values.
- In short:
  - D_i directly summarizes how much *all* of the fitted (predicted) values change when the i^{th} observation is deleted.
  - **A data point having a large D_i indicates that the data point strongly influences the fitted values (i.e., strongly influences model conclusions)**

Outlier Diagnostic Measures – Cook's D(istance)

- **Cook's Distance (D)**
- $\mathbf{D_i} = \frac{e_i^2}{(k+1)*MSE}\left[\frac{h_i}{(1-h_i)^2}\right] = \frac{t_{res}^2}{(k+1)}\left[\frac{h_i}{(1-h_i)}\right]$ where k is the number of predictors
- Notice that Cook's D incorporates Measures of Leverage as Well as Outlierness (using studentized residuals)
- Here are the guidelines commonly used:
  - If D_i is greater than 0.5, then the i^{th} data point is worthy of further investigation as it **may be influential**.
  - If D_i is greater than 1, then the i^{th} data point is **quite likely to be influential**.
  - Or, if D_i sticks out like a sore thumb from the other D_i values, it is **almost certainly influential**.
- For observation 27, the Cook's D is 22.6! which makes it almost certainly influential

Handling Outliers

https://onlinecourses.science.psu.edu/stat462/node/174

- Is an outlier "meaningful" or the result of an error?
- First, check for obvious data errors:
  - If the error is just a data entry or data collection error, correct it.
  - If the data point is not representative of the intended study population, delete it.
  - If the data point is a procedural error and invalidates the measurement, delete it.
- Consider the possibility that you might have just misformulated your regression model:
  - Did you leave out any important predictors?
  - Should you consider adding some interaction terms?
  - Is there any nonlinearity that needs to be modeled? (**our situation**)
- Decide whether or not deleting data points is warranted:
  - Do not delete data points just because they do not fit your preconceived regression model.
  - You must have a good, **objective** reason for deleting data points.
  - If you delete any data after you've collected it, justify and describe it in your reports.
  - If you are not sure what to do about a data point, analyze the data twice — once with and once without the data point — and report the results of both analyses.
- First, foremost, and finally — it's okay to use your common sense and knowledge about the situation.

Handling Outliers

- In our case, we see that the log-transformation of Home also transformed what appeared to be influential outliers
- We can see that after transforming the Home variable, none of the observations we considered earlier are influential outliers.
- Of course, we should run the full model with all important predictors and re-check the observations

	logHome	assets	leverage	stud_res	cook_dist
1	6.2120	142.49	0.0482	0.5037	0.0065
2	2.7856	117.29	0.0134	0.3821	0.0010
3	6.0584	134.15	0.0458	0.0699	0.0001
4	4.7287	139.08	0.0285	0.8931	0.0117
5	-2.6578	100.01	0.0357	1.5669	0.0448
6	-1.7714	85.72	0.0256	0.3483	0.0016
7	2.4493	106.20	0.0120	-0.1314	0.0001
8	-2.5245	64.02	0.0340	-0.6286	0.0070
9	3.2268	108.16	0.0157	-0.3256	0.0009
10	1.5933	97.62	0.0101	-0.2923	0.0004
11	1.9474	82.68	0.0106	-1.3134	0.0092
12	0.0393	87.33	0.0127	-0.2772	0.0005
13	1.0261	126.67	0.0102	1.6451	0.0136
14	1.2726	87.73	0.0100	-0.7430	0.0028
15	0.0678	72.81	0.0126	-1.1419	0.0083
16	-3.2164	43.93	0.0434	-1.5607	0.0544
17	3.0722	83.81	0.0148	-1.7096	0.0216
18	3.8031	143.63	0.0198	1.5376	0.0235
19	1.6620	135.06	0.0102	1.8935	0.0179
20	-1.6602	95.27	0.0245	0.8668	0.0094
21	-2.4068	58.77	0.0326	-0.9879	0.0164
22	0.1989	69.31	0.0121	-1.4039	0.0119
23	3.2125	110.59	0.0156	-0.1780	0.0003
24	-1.3859	82.75	0.0219	0.0203	0.0000
25	-0.0618	98.88	0.0132	0.4365	0.0013
26	-0.1391	115.88	0.0135	1.4740	0.0147
27	7.6456	129.77	0.0740	-0.8443	0.0286
28	3.7060	112.83	0.0190	-0.2433	0.0006
29	-1.6089	109.37	0.0240	1.6936	0.0346
30	3.3113	131.68	0.0163	1.0194	0.0086
31	1.6390	115.40	0.0101	0.7264	0.0027
32	5.3467	150.13	0.0358	1.3082	0.0316
33	4.1268	127.67	0.0225	0.4584	0.0024
34	1.2613	85.36	0.0100	-0.8778	0.0039
35	2.2597	114.55	0.0114	0.4303	0.0011
36	0.8587	44.45	0.0104	-3.2617	0.0507
37	-9.2103	46.52	0.1889	1.1171	0.1450
38	-0.7338	118.50	0.0169	1.8826	0.0297
39	0.5008	98.91	0.0111	0.2149	0.0003
40	1.5390	84.44	0.0101	-1.0435	0.0055
41	0.6831	87.24	0.0107	-0.5374	0.0016
42	-1.6602	60.45	0.0245	-1.1862	0.0176
43	0.9361	125.20	0.0103	1.5931	0.0130
44	-0.0407	87.77	0.0131	-0.2198	0.0003
45	-3.5032	50.85	0.0477	-1.0243	0.0263
46	-0.1277	99.73	0.0135	0.5125	0.0018
47	5.2254	116.79	0.0343	-0.6212	0.0069
48	2.3273	96.44	0.0116	-0.6529	0.0025
49	0.9322	94.40	0.0103	-0.2183	0.0002
50	2.3155	124.66	0.0115	1.0018	0.0059
51	3.8501	127.85	0.0201	0.5788	0.0035
52	-2.4068	90.48	0.0326	0.8872	0.0133
53	2.5112	103.38	0.0122	-0.3203	0.0006
54	2.1066	112.67	0.0109	0.3812	0.0008
55	6.4087	118.41	0.0513	-1.0123	0.0277
56	1.9685	130.82	0.0106	1.5093	0.0121

HANDLING MISSING DATA

Lecture 4c-Part 6

Missing Data

- In practice, it is common to find records with data missing on one or more variables.
- How missing data is handled depends on:
  - the pattern (or lack of pattern) in the missing data
  - the type of analysis you are doing
  - the type of variable(s) for which values are missing, and
  - whether the variables with missing data are involved in the analysis.
- We will look at basic ways of handling missing data; some techniques are advanced and beyond the scope of this course
- Many software (as well as the R language) have special capabilities for analyzing data with missing values.

Missing Data - Pattern

- The first thing we have to establish is the pattern (or lack of pattern) in the missing data. This determines whether the results of the analysis are biased or unbiased. Two **ignorable** patterns are

- **MCAR** (Missing Completely at Random) –
  - Presence or absence of data in X is unrelated to the observed data on other variables, as well as the variable with missing data. In other words, the missing data are just a random subset of the data
    - Example: Missing values of Assets do not depend on Gender on or the value of Asset (Males and Females equally likely to have missing values, and we do not systematically miss Assets for low or high values)

- **MAR** (Missing at Random) –
  - Presence or absence of data in X is related to the observed data of other variables but not to the missing data.
    - Example: Missing values of Assets depend on Gender (say Males more likely to have missing values) but NOT on the value of Assets (we do not systematically miss Assets for low or high values)

- **Ignorable** data loss and small percentages of missing values usually not a cause for concern

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	47.4	
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	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		34.4	
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		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		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		110.59
24	0	0	60	0.25	3.95	82.75
25	0	1		0.94	43.36	98.88
26	0	1	47	0.87	83.12	115.88
27	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

Missing Data - Pattern

- **Non-ignorable** or **Systematic** pattern –
  - Also called Missing NOT at Random (MNAR)
  - Missing data in X (say, Assets) depends on other variables (such as Gender) as well as the values of X (Assets) itself.
  - For example, Males with Lower assets are more likely to have missing values.
  - If the analysis involves Assets and Gender, then the results of the analysis can be seriously biased.

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	47.4	
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	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		34.4	
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		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		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		110.59
24	0	0	60	0.25	3.95	82.75
25	0	1		0.94	43.36	98.88
26	0	1	47	0.87	83.12	115.88
27	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

Missing Data - Pattern

- **Detecting** whether missing data is Ignorable (MAR/MCAR) or Systematic (Missing Not at Random – MNAR)
  - No single test – examine various features of the data set
  - MAR/MCAR tests – use ANOVA/Regression to compare observations with values missing on some variables with observations without missing values. Significant differences may point to data loss mechanism
  - Beyond the scope of this course

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	
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	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		34.49	
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		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		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		110.59
24	0	0	60	0.25	3.95	82.75
25	0	1		0.94	43.36	98.88
26	0	1	47	0.87	83.12	115.88
27	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

Missing Data - Handling

- **Two Basic categories**
  - Delete Missing Data (Listwise vs Pairwisewise)
  - Replace Missing data with Other values (Simple Imputation Methods)
- Regardless of method of handling it is recommended that you analyze data with different approaches and compare the results.
- Explain in written summaries the extent of missing data and the steps taken

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	47.4	
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	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		34.4	
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		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		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		110.59
24	0	0	60	0.25	3.95	82.75
25	0	1		0.94	43.36	98.88
26	0	1	47	0.87	83.12	115.88
27	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

Missing Data - Handling

- **Delete Missing Data**
  - Assumes MCAR or MAR
  - Does not take advantage of information in data
  - Often yield biased results under the less strict assumption of MAR.
  - More biased results when loss is systematic

- **Listwise deletion**:
  - Cases with missing scores *on any variable* are excluded from all analyses; the effective sample includes only cases with complete records.
  - Standard errors estimated after listwise deletion usually larger
  - In listwise deletion (unlike pairwise) all analyses are conducted with the same number of cases.

- **Pairwise deletion**:
  - Observations are excluded *only if they have missing data on variables involved in a particular analysis*. For example, assume that the only missing value is for gender in Observation 6.
  - Then, delete observation 6 when the model involves Gender (say a model with Gender, Mortgage Balance and Assets). However, if the analysis does not include Gender (say, you are analyzing Assets vs Mortgage Balance) the values of Assets and Mortgage Balance for Observation 6 will still be used.
  - You can see that when Gender is used in the analysis, the sample size is different than when Gender is not used in the analysis.
  - Also note that the correlation between Assets and Mortgage Balance depends on whether Gender is included in the analysis or not. This can create problems in the analysis as well as bias conclusions.

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	47.4	
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	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		34.4	
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		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		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		110.59
24	0	0	60	0.25	3.95	82.75
25	0	1		0.94	43.36	98.88
26	0	1	47	0.87	83.12	115.88
27	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

Missing Data - Handling

- **Simple Imputation (when you don’t delete observations)**:
    - All methods tend to underestimate error variance, especially for large number of missing cases
    - **Mean Substitution**:
        - Replace missing values with mean or group mean if group membership is a predictor in the analysis.
        - Disadvantage is reduced variance which can distort results.
        - Also makes distribution more peaked at the mean and distorts it.
    - **Regression-based imputation**:
        - Missing score replaced by predicted score using multiple regression based on available scores on other variables.
        - Uses more information than mean substitution
        - Use entire sample, and not just data from one group, to avoid range restriction caused by

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	47.4	
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	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		34.4	
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		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		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		110.59
24	0	0	60	0.25	3.95	82.75
25	0	1		0.94	43.36	98.88
26	0	1	47	0.87	83.12	115.88
27	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

Missing Data - Handling

- **Hot Deck Imputation**:
    - Separate complete from incomplete records
    - Sort both sets (decks) so that cases with similar profiles on other variables are grouped together
    - For each record with a missing value on a variable, we will now have a collection of records which are similar (on the other variables)
  - **Deterministic Imputation**:
    - Replace a missing value with a score from an observation with the most similar profile of scores in other variables (i.e., nearest neighbor).
  - **Random Imputation**:
    - Replace missing scores with those on the same variable from a random selection of similar complete records.
- Hot Deck Imputation generally result in less loss of information and distortion relative to other methods such as mean substitution.
- More sophisticated Imputation Methods also available but beyond our scope

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	47.4	
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	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		34.4	
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		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		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		110.59
24	0	0	60	0.25	3.95	82.75
25	0	1		0.94	43.36	98.88
26	0	1	47	0.87	83.12	115.88
27	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