جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Regression Analysis

Regression Diagnostics

Spring 2012

1 The Multiple Regression Model

Multiple Regression Analysis

Motivation

Example (Where to locate new outlets)

Characteristics of the community that surrounds a site, such as the size and affluence of the local population, influence the success of a new restaurant.

Multiple Regression Analysis

Motivation

Example (Where to locate new outlets)

Characteristics of the community that surrounds a site, such as the size and affluence of the local population, influence the success of a new restaurant. It would not make sense to locate an expensive restaurant that caters to small business meals in a neighborhood of large, working-class families.

Multiple Regression Analysis

Motivation

Example (Where to locate new outlets)

Characteristics of the community that surrounds a site, such as the size and affluence of the local population, influence the success of a new restaurant. It would not make sense to locate an expensive restaurant that caters to small business meals in a neighborhood of large, working-class families.
A location that looks good to one business, however, most likely appeals to competitors as well.

Multiple Regression Analysis

Motivation

Example (Where to locate new outlets)

Characteristics of the community that surrounds a site, such as the size and affluence of the local population, influence the success of a new restaurant. It would not make sense to locate an expensive restaurant that caters to small business meals in a neighborhood of large, working-class families.

A location that looks good to one business, however, most likely appeals to competitors as well.

Locations near an upscale shopping mall are likely to have many competitors as well.

Which is better: to be far from the competition or to be in a more affluent area?

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Multiple Regression Analysis

Motivation

Example (Where to locate a new outlet)

A chain is considering where to locate a new restaurant. Is it better to locate it far from the competition or in a more affluent area?

Multiple Regression Analysis

Motivation

Example (Where to locate a new outlet)

A chain is considering where to locate a new restaurant. Is it better to locate it far from the competition or in a more affluent area?

- Use multiple regression to describe the relationship between several explanatory variables and the response.

Multiple Regression Analysis

Motivation

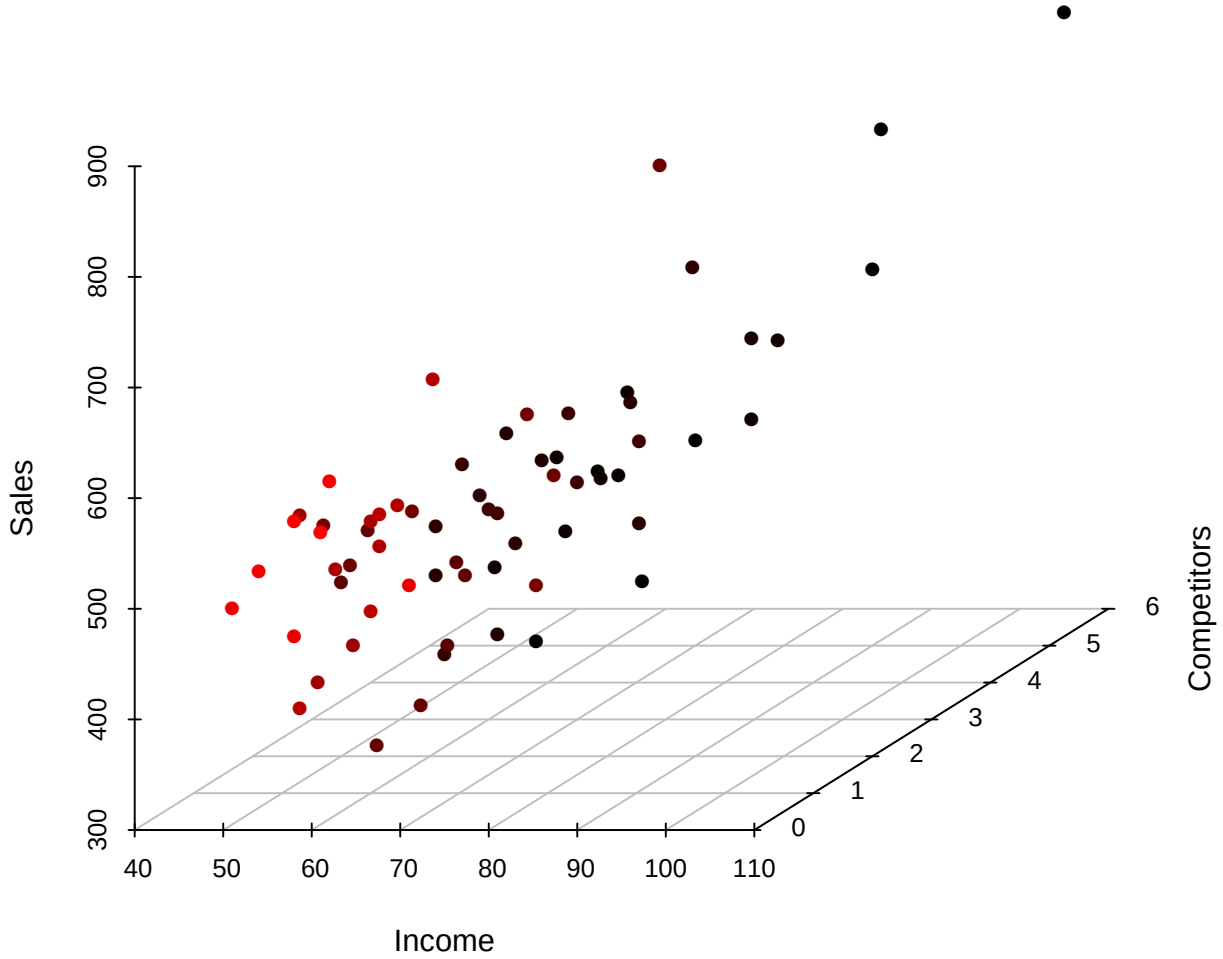
Example (Where to locate a new outlet)

A chain is considering where to locate a new restaurant. Is it better to locate it far from the competition or in a more affluent area?

- Use multiple regression to describe the relationship between several explanatory variables and the response.
- Multiple regression separates the effects of each explanatory variable on the response and reveals which really matter

Multiple Regression Analysis

Example: where to locate a new outlet



جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Multiple Regression Analysis

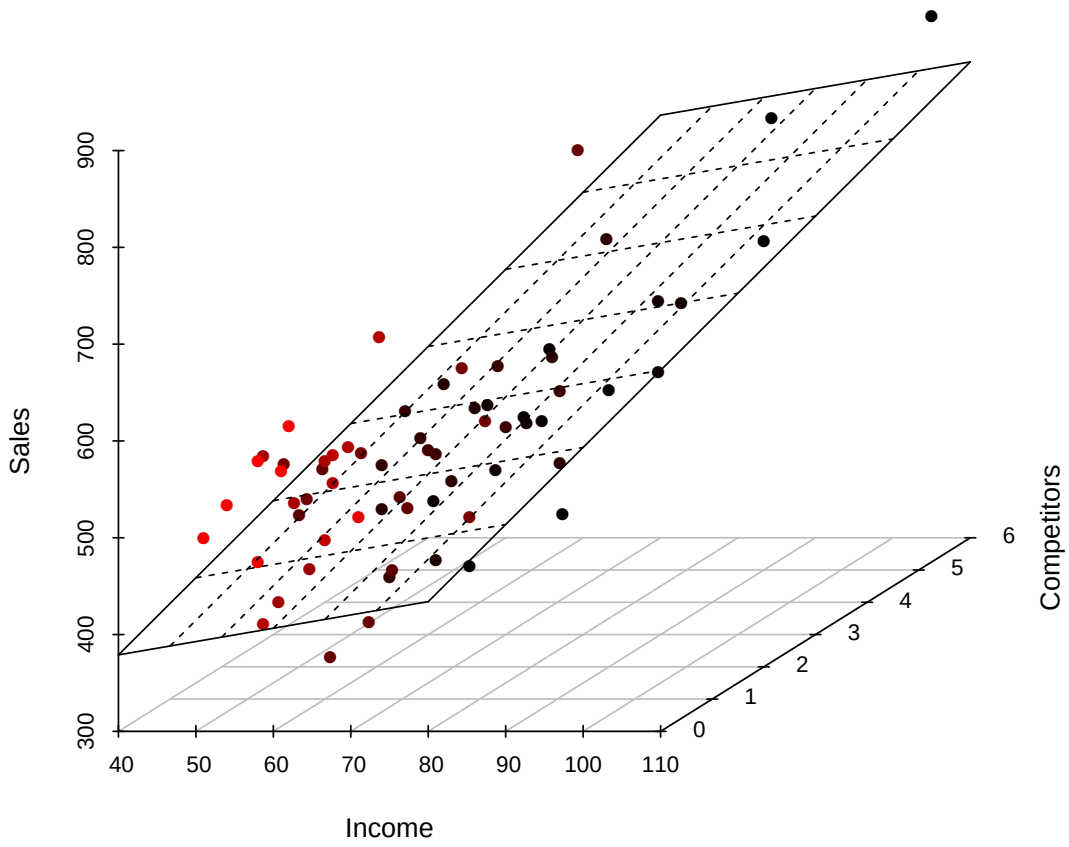
Example: where to locate a new outlet

$$\texttt{Sales}_i = \beta_0 + \beta_1 \cdot \texttt{Income}_i + \beta_2 \cdot \texttt{Competitors}_i + \epsilon_i$$

Multiple Regression Analysis

Example: where to locate a new outlet

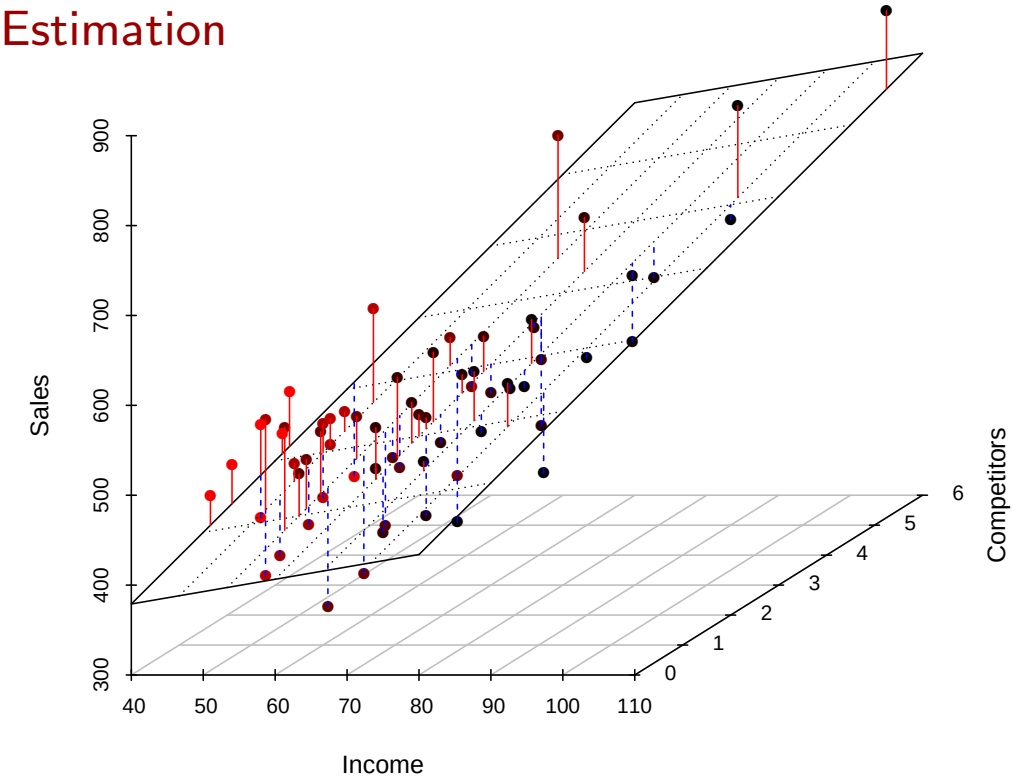
$$\texttt{Sales}_i = \beta_0 + \beta_1 \cdot \texttt{Income}_i + \beta_2 \cdot \texttt{Competitors}_i + \epsilon_i$$



جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Multiple Regression Analysis

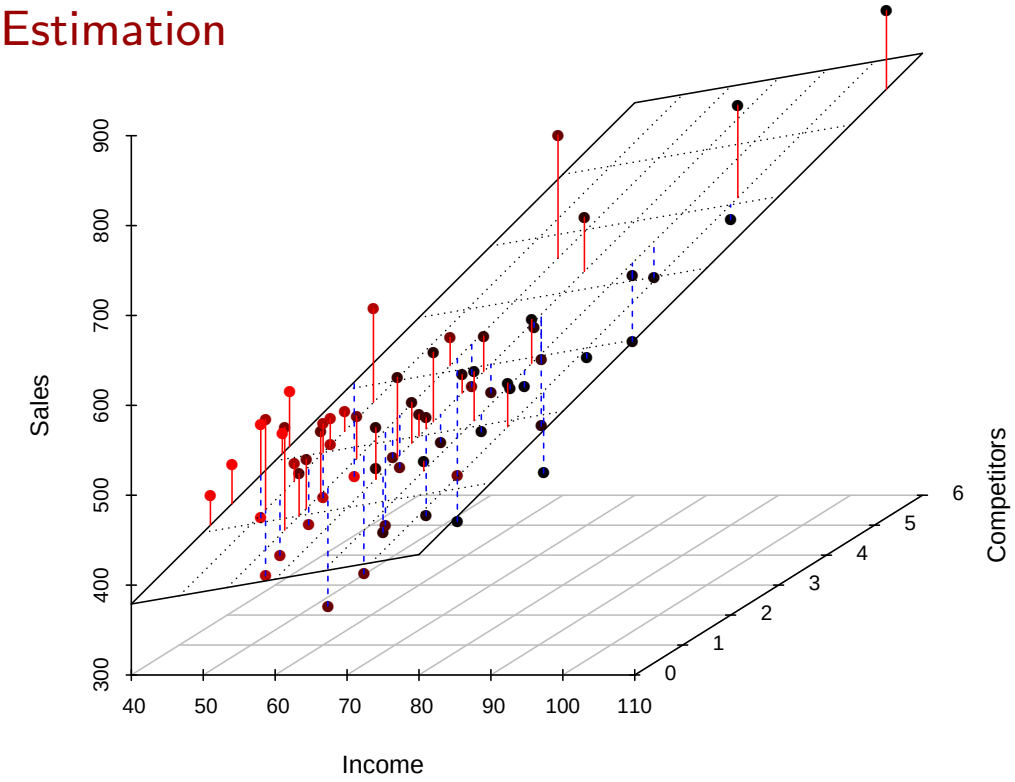
Least Squares Estimation



جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Multiple Regression Analysis

Least Squares Estimation



The unobserved errors in the model

- are independent of one another,
- have equal variance, and
- are normally distributed around the regression equation.

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Multiple Regression Analysis

- While the SRM bundles all but one explanatory variable into the error term, multiple regression allows for the inclusion of several variables in the model.
- In the MRM, residuals departing from normality may suggest that an important explanatory variable has been omitted.

Multiple Regression Analysis

- While the SRM bundles all but one explanatory variable into the error term, multiple regression allows for the inclusion of several variables in the model.
- In the MRM, residuals departing from normality may suggest that an important explanatory variable has been omitted.

Example (Women's Apparel Stores)

Response variable: sales at stores in a chain of women's apparel (annually in dollars per square foot of retail space).

Two explanatory variables: median household income in the area (thousands of dollars) and number of competing apparel stores in the same mall.

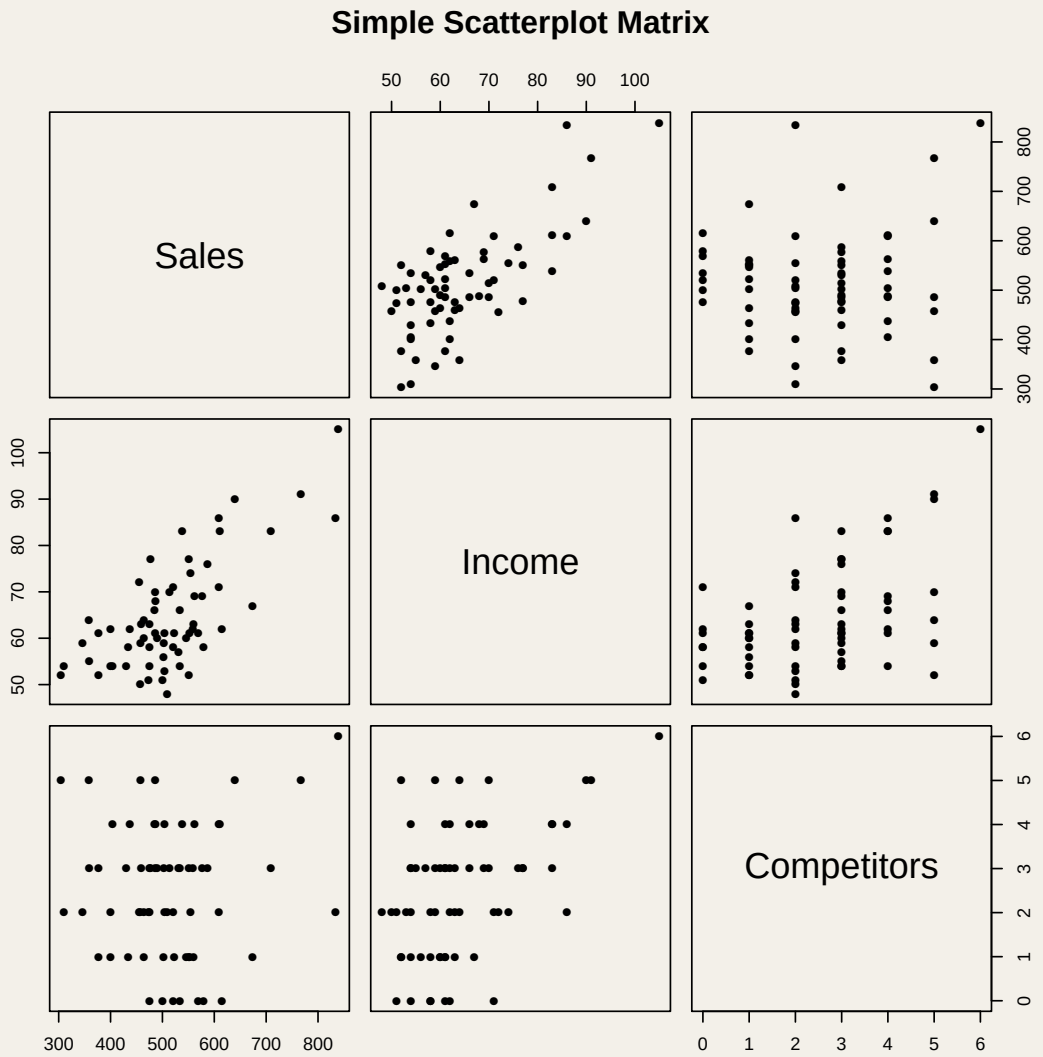
جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Example (Women's Apparel Stores)

- Begin with a scatterplot matrix, a table of scatterplots arranged as in a correlation matrix.
- Using a scatterplot matrix to understand data can save considerable time later when interpreting the multiple regression results.

Multiple Regression Analysis

Example (Women's Apparel Stores)



جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Multiple Regression Analysis

Example (Example: Women's Apparel Stores)

The scatterplot matrix for this example

- Confirms a positive linear association between sales and median household income.
- Shows a weak association between sales and number of competitors.

Multiple Regression Analysis

Example (Example: Women's Apparel Stores)

The scatterplot matrix for this example

- Confirms a positive linear association between sales and median household income.
- Shows a weak association between sales and number of competitors.

Correlation Matrix

	Sales	Income	Competitors
Sales	1.0000	0.7080	0.0666
Income	0.7080	1.0000	0.4743
Competitors	0.0666	0.4743	1.0000

Multiple Regression Analysis

Software Output

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.3587    49.2902   1.225 0.225374
Income        7.9660     0.8382   9.503 0.000000
Competitors -24.1650     6.3899  -3.782 0.000000

Residual standard error: 68.03 on 62 df
Multiple R-squared: 0.5947, Adjusted R-squared: 0.5817
F-statistic: 45.49 on 2 and 62 DF,  p-value: 0.0000
```


جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Multiple Regression Analysis

R^2 and s_e.

- R^2 indicates that the fitted equation explains 59.47% of the store-to-store variation in sales.
- For this example, R^2 is larger than the r^2 values for separate SRMs fitted for each explanatory variable; it is also larger than their sum.
- For this example, se $= \$68.03$.
- $\bar{R}^2$ is known as the adjusted R-squared. It adjusts for both sample size n and model size k. It is always smaller than R^2.
- The residual degrees of freedom $(n - k - 1)$ is the divisor of s_e. $\bar{R}^2$ and s_e move in opposite directions when an explanatory variable is added to the model ($\bar{R}^2$ goes up while s_e goes down).

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Multiple Regression Analysis

Marginal and Partial Slopes

- Partial slope: slope of an explanatory variable in a multiple regression that statistically excludes the effects of other explanatory variables.
- Marginal slope: slope of an explanatory variable in a simple regression.

Multiple Regression Analysis

Marginal and Partial Slopes

- Partial slope: slope of an explanatory variable in a multiple regression that statistically excludes the effects of other explanatory variables.
- Marginal slope: slope of an explanatory variable in a simple regression.

Partial Slopes: Women's Apparel Stores

- The slope $\hat{\beta}_1 = 7.966$ for Income implies that a store in a location with a higher median household of \$10,000 sells, on average, \$79.66 more per square foot than a store in a less affluent location with the same number of competitors.

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Multiple Regression Analysis

Marginal and Partial Slopes

- Partial slope: slope of an explanatory variable in a multiple regression that statistically excludes the effects of other explanatory variables.
- Marginal slope: slope of an explanatory variable in a simple regression.

Partial Slopes: Women's Apparel Stores

- The slope $\hat{\beta}_1 = 7.966$ for Income implies that a store in a location with a higher median household of \$10,000 sells, on average, \$79.66 more per square foot than a store in a less affluent location with the same number of competitors.
- The slope $\hat{\beta}_2 = -24.165$ implies that, among stores in equally affluent locations, each additional competitor lowers average sales by \$24.165 per square foot.

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Multiple Regression Analysis

Marginal and Partial Slopes

- Partial and marginal slopes only agree when the explanatory variables are uncorrelated.
- In this example they do not agree. For instance, the marginal slope for Competitors is 4.6352. It is positive because more affluent locations tend to draw more competitors. The MRM separates these effects but the SRM does not.

Checking Conditions

Conditions for inference

Use the residuals from the fitted MRM to check that the errors in the model

- are independent;
- have equal variance; and
- follow a normal distribution.

Checking Conditions

Conditions for inference

Use the residuals from the fitted MRM to check that the errors in the model

- are independent;
- have equal variance; and
- follow a normal distribution.

Two scatterplots summarize the overall fit of a multiple regression. These plots are analogous to the scatterplot of Y on X and the scatterplot of e on X used in simple regression. Indeed, most diagnostic plots used to check a multiple regression convert it into a simple regression in one way or another.

Checking Conditions

Calibration Plot

A calibration plot is analogous to the scatterplot of Y on X.

- Calibration plot: scatterplot of the response Y on the fitted values $\hat{Y}$.
- R^2 is the squared correlation between Y and $\hat{Y}$; the tighter data cluster along the diagonal line in the calibration plot, the larger the R^2 value.

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Checking Conditions

Calibration Plot

A calibration plot is analogous to the scatterplot of Y on X.

- Calibration plot: scatterplot of the response Y on the fitted values $\hat{Y}$.
- R^2 is the squared correlation between Y and $\hat{Y}$; the tighter data cluster along the diagonal line in the calibration plot, the larger the R^2 value.

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Checking Conditions

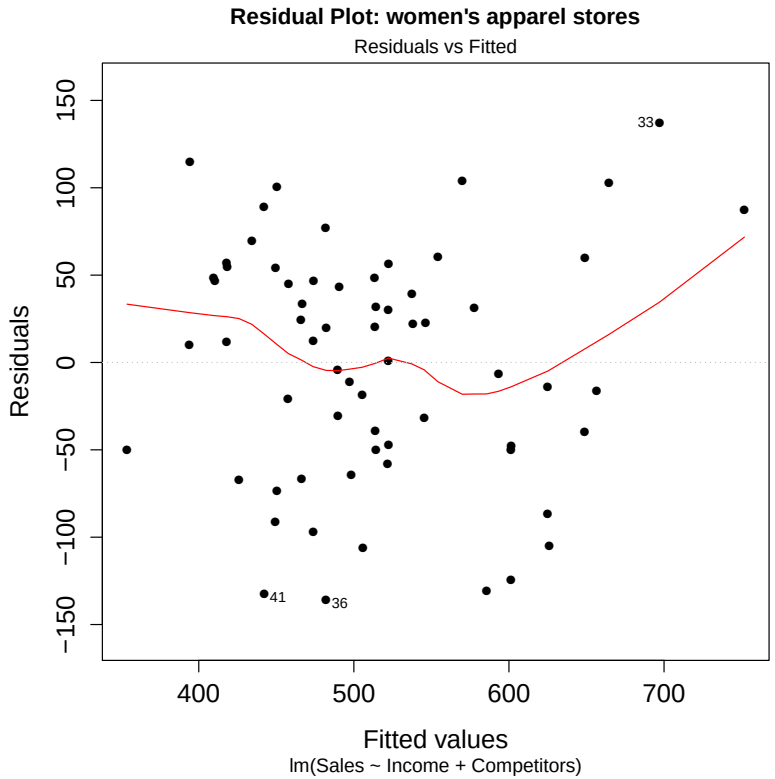
Residual Plot

- Plot of residuals versus fitted Y values is used to identify outliers and to check for the similar variances condition.
- Plot of residuals versus each explanatory variable are used to verify that the relationships are linear.

Checking Conditions

Residual Plot

- Plot of residuals versus fitted Y values is used to identify outliers and to check for the similar variances condition.
- Plot of residuals versus each explanatory variable are used to verify that the relationships are linear.



Checking Conditions

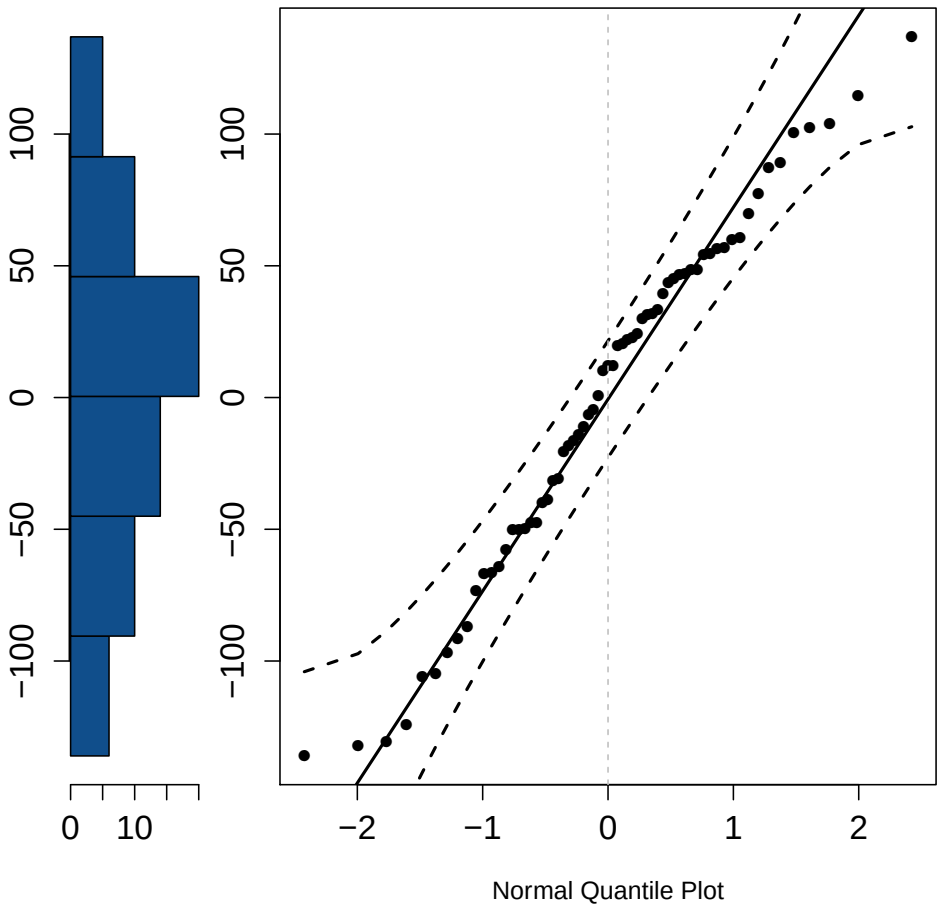
Residual Plot

These plot of residuals versus Income and Competitors have no evident pattern.

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Checking Conditions

Check Normality: WomenÕs Apparel Stores



The quantile plot indicates nearly normal condition is satisfied.

Inference in Multiple Regression

Inference for the Model: F-test

- F-test: test of the explanatory power of the MRM as a whole.

$$H_0 \quad : \quad \beta_1 = \beta_2 = \ldots = \beta_K = 0$$

$$H_1 \quad : \quad \texttt{At least one coefficient is not equal to zero}$$

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Inference in Multiple Regression

Inference for the Model: F-test

- F-test: test of the explanatory power of the MRM as a whole.

$$H_0 \quad : \quad \beta_1 = \beta_2 = \ldots = \beta_K = 0$$

$$H_1 \quad : \quad \texttt{At least one coefficient is not equal to zero}$$

- F-statistic: ratio of the sample variance of the fitted values to the variance of the residuals.

Inference in Multiple Regression

Inference for the Model: F-test

- F-test: test of the explanatory power of the MRM as a whole.

$$H_0 \quad : \quad \beta_1 = \beta_2 = \ldots = \beta_K = 0$$
$$H_1 \quad : \quad \texttt{At least one coefficient is not equal to zero}$$

- F-statistic: ratio of the sample variance of the fitted values to the variance of the residuals.

$$F = \frac{\texttt{MSR}}{\texttt{MSE}}$$

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Inference in Multiple Regression

Inference for the Model: F-test

- F-test: test of the explanatory power of the MRM as a whole.

$$H_0 \quad : \quad \beta_1 = \beta_2 = \ldots = \beta_K = 0$$

$$H_1 \quad : \quad \texttt{At least one coefficient is not equal to zero}$$

- F-statistic: ratio of the sample variance of the fitted values to the variance of the residuals.

$$F = \frac{\texttt{MSR}}{\texttt{MSE}}$$

$$F \sim F_{k,n-k-1}$$

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Inference in Multiple Regression

F-test Results in Analysis of Variance Table

The `p-value` for the $F-$statistics is typically located in the analysis of variance table, or ANOVA table. This table summarizes the overall fit of the regression, i.e., it gives a detailed accounting of the variation in Y.

Inference in Multiple Regression

F-test Results in Analysis of Variance Table

The `p-value` for the $F-$statistics is typically located in the analysis of variance table, or ANOVA table. This table summarizes the overall fit of the regression, i.e., it gives a detailed accounting of the variation in Y.

Example (ANOVA Table: Women's Apparel Stores)

	DF	SS	MS	F	P
Regression	2	421107	210554	45.494	0.0000
Residual Error	62	286946	4628		
Total	64	708053			

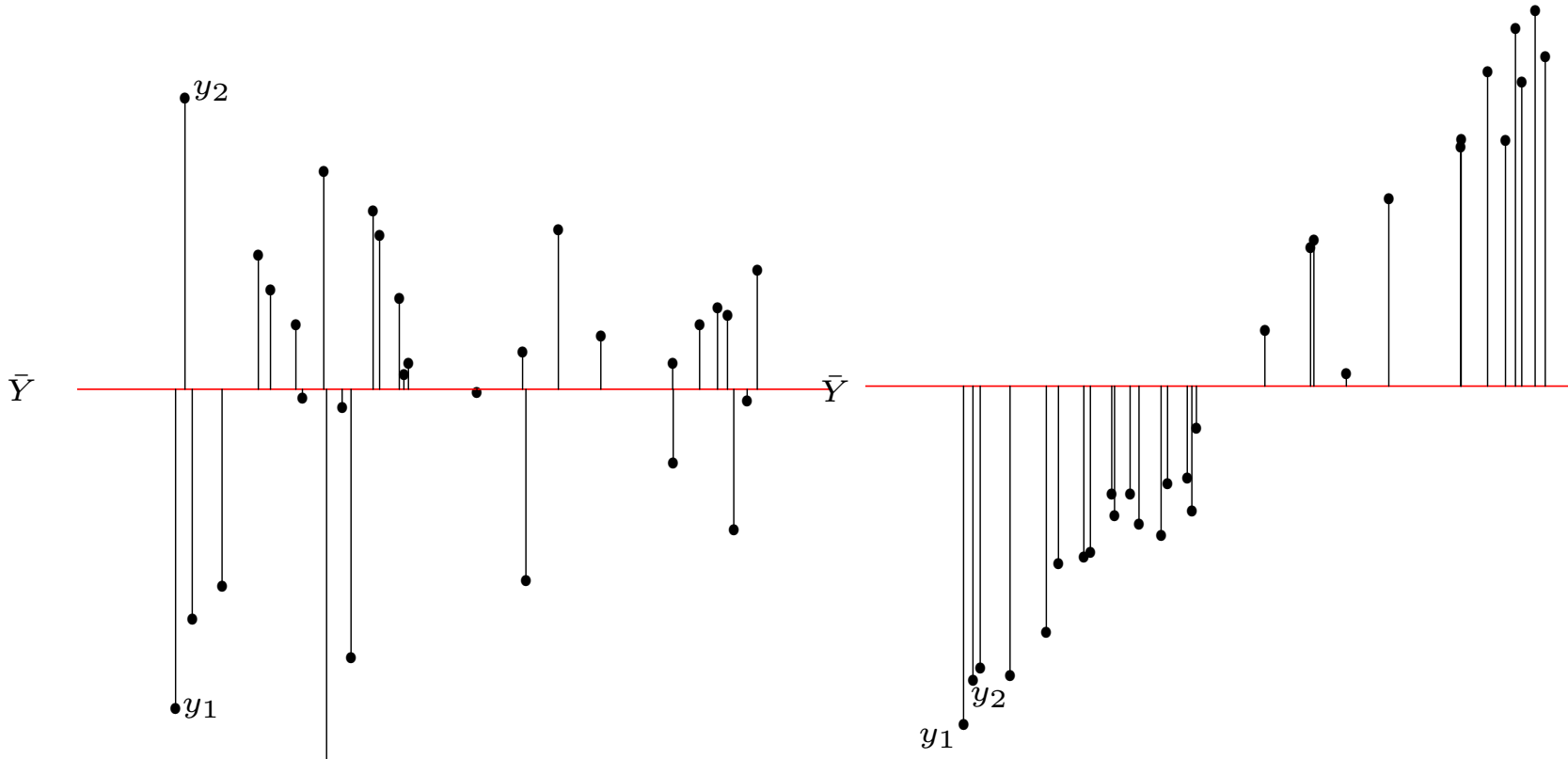
The F-statistic has a `p-value` of < 0.0001; reject H_0. Income and Competitors together explain statistically significant variation in sales.

Inference in Multiple Regression

F-test Results in Analysis of Variance Table

Total Sum-of-Squares

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$



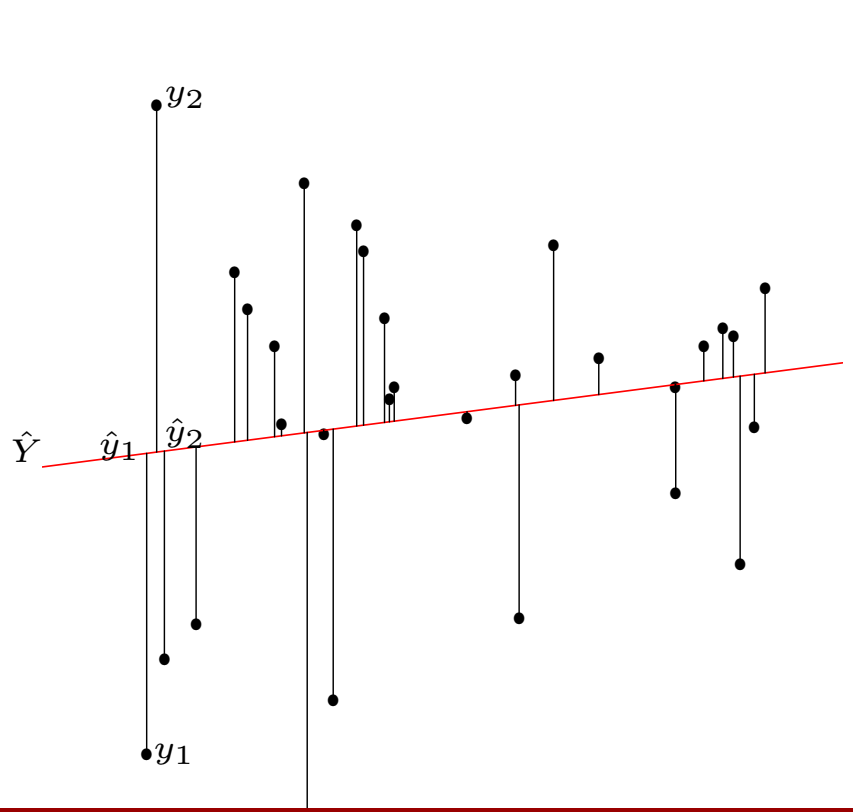
جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

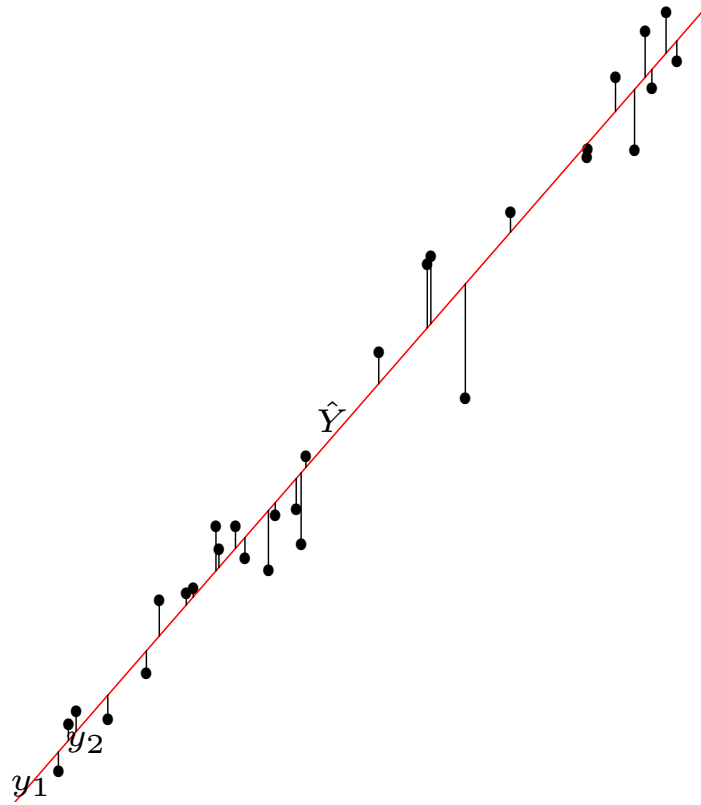
Inference in Multiple Regression

F-test Results in Analysis of Variance Table

Error Sum-of-Squares

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





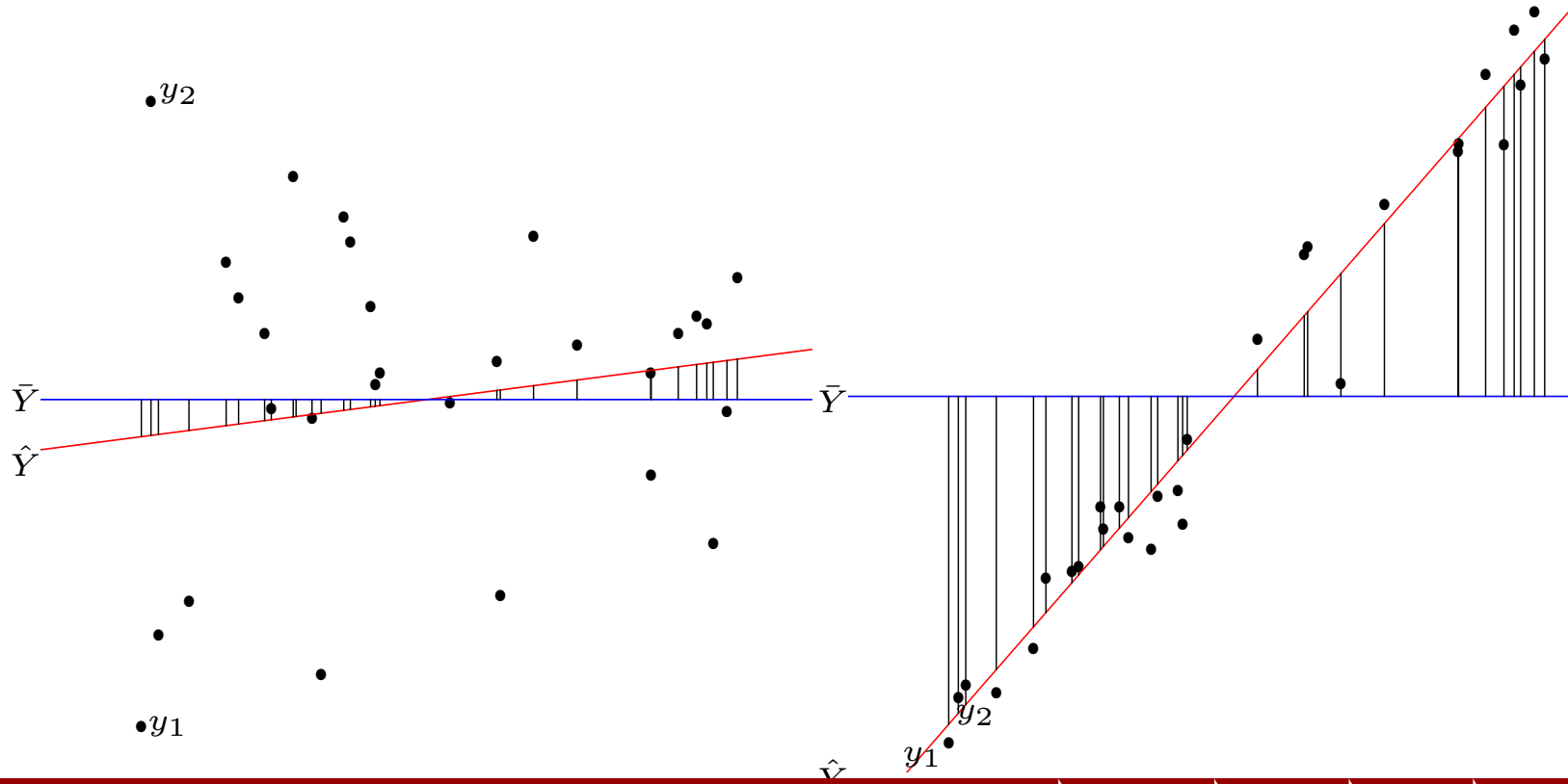
جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Inference in Multiple Regression

F-test Results in Analysis of Variance Table

Regression Sum-of-Squares

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$



جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Inference in Multiple Regression

R^2 and R^2_{adj}

The sum-of-squares: SST, SSR and SSE can be used to evaluate how well the regression equation is explaining the variation of Y. One measure of the goodness of fit of the regression is the coefficient of determination:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

Inference in Multiple Regression

R^2 and R^2_{adj}

The sum-of-squares: SST, SSR and SSE can be used to evaluate how well the regression equation is explaining the variation of Y. One measure of the goodness of fit of the regression is the coefficient of determination:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

R^2 represents the proportion of the variation in Y explained by the regression. Because the addition of **ANY** explanatory variables to the model causes SSE to decrease, R^2 gets increasingly closer than 1. We need an adjusted version of R^2:

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Inference in Multiple Regression

Inference for one coefficient

The t-statistic is used to test each slope using the null hypothesis $H_0 : \beta_j = 0$.

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Inference in Multiple Regression

Inference for one coefficient

The t-statistic is used to test each slope using the null hypothesis $H_0 : \beta_j = 0$.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   60.3587    49.2902   1.225 0.225374
Income         7.9660     0.8382   9.503 0.000000
Competitors  -24.1650     6.3899  -3.782 0.000000
```


جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Inference in Multiple Regression

Inference for one coefficient

The t-statistic is used to test each slope using the null hypothesis $H_0 : \beta_j = 0$.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   60.3587    49.2902   1.225 0.225374
Income         7.9660     0.8382   9.503 0.000000
Competitors  -24.1650     6.3899  -3.782 0.000000
```

The t-statistics and associated `p-value` indicate that both slopes are significantly different from zero.

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Inference in Multiple Regression

Inference for one coefficient

Hypotheses

$$H_0 : \beta_0 = \beta_0^* \qquad H_0 : \beta_1 = \beta_1^* \qquad H_0 : \beta_2 = \beta_2^*$$
$$H_1 : \beta_0 \neq \beta_0^* \qquad H_1 : \beta_1 \neq \beta_1^* \qquad H_1 : \beta_2 \neq \beta_2^*$$

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Inference in Multiple Regression

Inference for one coefficient

Hypotheses

$$H_0 : \beta_0 = \beta_0^* \qquad H_0 : \beta_1 = \beta_1^* \qquad H_0 : \beta_2 = \beta_2^*$$
$$H_1 : \beta_0 \neq \beta_0^* \qquad H_1 : \beta_1 \neq \beta_1^* \qquad H_1 : \beta_2 \neq \beta_2^*$$

Test Statistic

$$t = \frac{\beta_0 - \beta_0^*}{S_{\hat{\beta}_0}} \qquad t = \frac{\beta_1 - \beta_1^*}{S_{\hat{\beta}_1}} \qquad t = \frac{\beta_2 - \beta_2^*}{S_{\hat{\beta}_2}}$$

Inference in Multiple Regression

Inference for one coefficient

Hypotheses

$$H_0 : \beta_0 = \beta_0^* \qquad H_0 : \beta_1 = \beta_1^* \qquad H_0 : \beta_2 = \beta_2^*$$
$$H_1 : \beta_0 \neq \beta_0^* \qquad H_1 : \beta_1 \neq \beta_1^* \qquad H_1 : \beta_2 \neq \beta_2^*$$

Test Statistic

$$t = \frac{\beta_0 - \beta_0^*}{S_{\hat{\beta}_0}} \qquad t = \frac{\beta_1 - \beta_1^*}{S_{\hat{\beta}_1}} \qquad t = \frac{\beta_2 - \beta_2^*}{S_{\hat{\beta}_2}}$$

Critical value

$$t_{n-k-1,1-\alpha/2}$$

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Inference in Multiple Regression

Inference for one coefficient

Confidence Intervals

$$\hat{\beta}_j \pm t_{n-k-1,1-\alpha/2} S_{\hat{\beta}_j}$$

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Inference in Multiple Regression

Inference for one coefficient

Confidence Intervals

$$\hat{\beta}_j \pm t_{n-k-1,1-\alpha/2} S_{\hat{\beta}_j}$$

Example (Confidence Intervals: Women's Apparel Stores)

$\hat{\beta}_0 \pm t_{62,.975} S_{\hat{\beta}_0}$	$\hat{\beta}_1 \pm t_{62,.975} S_{\hat{\beta}_1}$	$\hat{\beta}_2 \pm t_{62,.975} S_{\hat{\beta}_2}$
$60.359 \pm 1.999 \cdot 49.29$	$7.966 \pm 1.999 \cdot 0.838$	$-24.165 \pm 1.999 \cdot 6.39$
$(-38.171, 158.888)$	$(6.290, 9.642)$	$(-36.938, -11.392)$

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Inference in Multiple Regression

Prediction Intervals

- Let's get the predicted sales per square foot at a location with median income of \$70,000 and 3 competitors.

$\widehat{Sales} = 60.3587 + 7.9660 \cdot 70 - 24.165 \cdot 3 = \$545.48$ per square fo

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Inference in Multiple Regression

Prediction Intervals

- Let's get the predicted sales per square foot at a location with median income of \$70,000 and 3 competitors.

  $\widehat{Sales} = 60.3587 + 7.9660 \cdot 70 - 24.165 \cdot 3 = \$545.48$ per square fo

- From Minitab:

  95% CI: $(526.633, 564.331)$
  95% PI: $(408.191, 682.774)$

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Inference in Multiple Regression

Prediction Intervals

- Let's get the predicted sales per square foot at a location with median income of \$70,000 and 3 competitors.

  $\widehat{Sales} = 60.3587 + 7.9660 \cdot 70 - 24.165 \cdot 3 = \$545.48$ per square fo

- From Minitab:

  95% CI: $(526.633, 564.331)$

  95% PI: $(408.191, 682.774)$

- An approximate 95% prediction interval is given by $\hat{y} \pm 2s_e$.

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Inference in Multiple Regression

Prediction Intervals

- Let's get the predicted sales per square foot at a location with median income of \$70,000 and 3 competitors.

$$\widehat{Sales} = 60.3587 + 7.9660 \cdot 70 - 24.165 \cdot 3 = \$545.48 \text{ per square fo}$$

- From Minitab:

  95% CI: $(526.633, 564.331)$

  95% PI: $(408.191, 682.774)$

- An approximate 95% prediction interval is given by $\hat{y} \pm 2s_e$. For example, the 95% prediction interval for sales per square foot at a location with median income of \$70,000 and 3 competitors is approximately

$$\$545.48 \pm \$136.06 = (\$409.42, \$681.53) \text{ per square foot.}$$

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

Steps in Fitting a Multiple Regression (First version)

1. What is the problem to be solved? Do these data help in solving it?
2. Check the scatterplots of the response versus each explanatory variable (scatterplot matrix).
3. If the scatterplots appear straight enough, fit the multiple regression model. Otherwise find a transformation.
4. Obtain the residuals and fitted values from the regression.
5. Use residual plot of e vs. $\hat{y}$ to check for similar variance condition.
6. Construct residual plots of e vs. explanatory variables. Look for patterns.
7. Check whether the residuals are nearly normal.
8. Use the F-statistic to test the null hypothesis that the collection of explanatory variables has no effect on the response.
9. If the F-statistic is statistically significant, test and interpret individual partial slopes.

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

4M: Subprime Mortages

Motivation

A banking regulator would like to verify how lenders use credit scores to determine the interest rate paid by subprime borrowers. The regulator would like to separate its effect from other variables such as loan-to-value (LTV) ratio, income of the borrower and value of the home.

4M: Subprime Mortages

Motivation

A banking regulator would like to verify how lenders use credit scores to determine the interest rate paid by subprime borrowers. The regulator would like to separate its effect from other variables such as loan-to-value (LTV) ratio, income of the borrower and value of the home.

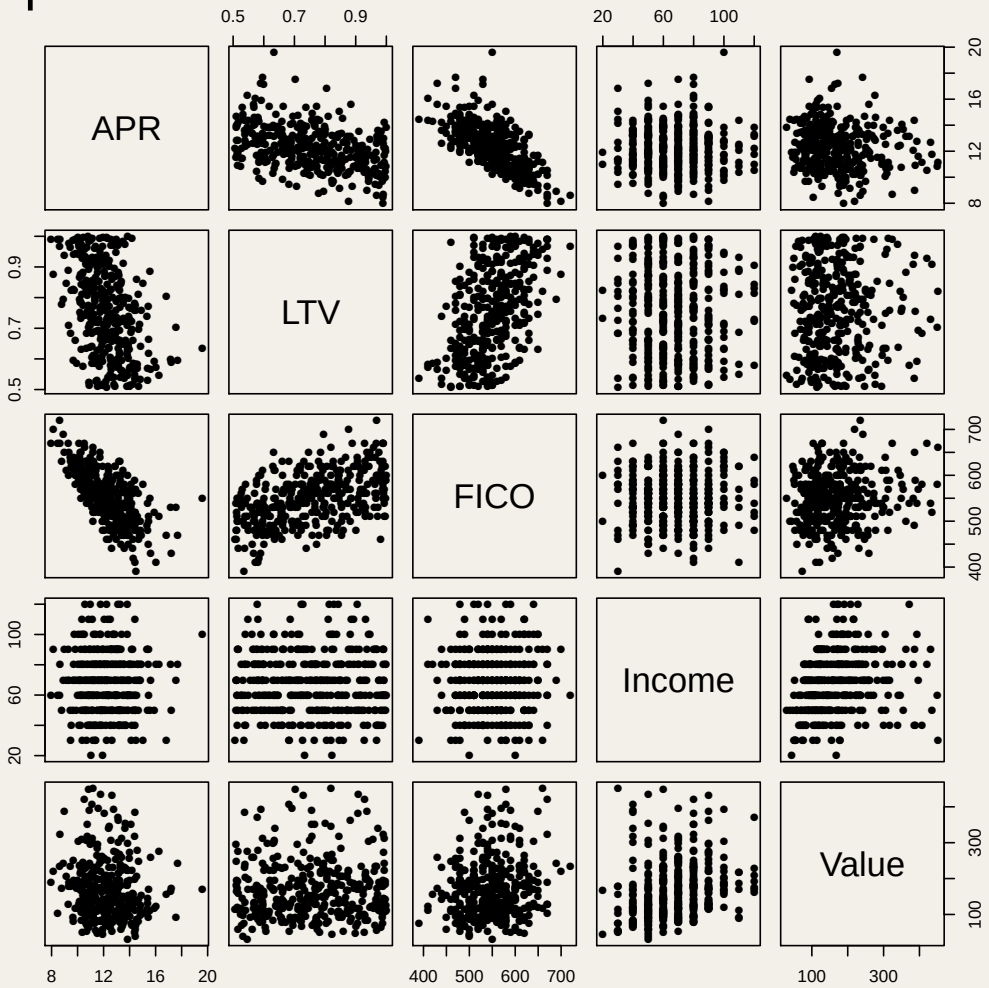
Method

Use multiple regression on data obtained for 372 mortgages from a credit bureau. The explanatory variables are the LTV, credit score, income of the borrower, and home value. The response is the annual percentage rate of interest on the loan (APR).

4M: Subprime Mortages

Method

Check scatterplot matrix



جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

4M: Subprime Mortages

Mechanics

Fit model and check conditions.

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

4M: Subprime Mortages

Mechanics

Fit model and check conditions.

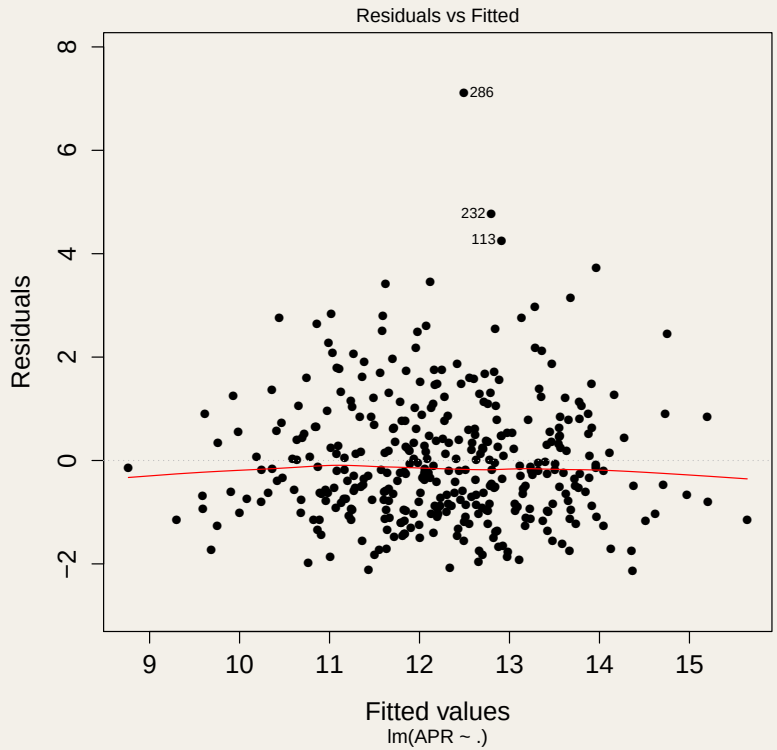
```
                Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.7253652  0.6859028  34.590   0.0000
LTV         -1.5888430  0.5197123  -3.057   0.0024
FICO        -0.0184318  0.0013502 -13.652   0.0000
Income       0.0004032  0.0033266   0.121   0.9036
Value       -0.0007521  0.0008186  -0.919   0.3589

Residual standard error: 1.244 on 367 degrees of freedo
Multiple R-squared: 0.4631, Adjusted R-squared: 0.4573
F-statistic: 79.14 on 4 and 367 DF,  p-value: < 2.2e-16
```

4M: Subprime Mortages

Mechanics

Fit model and check conditions. Residuals versus fitted values.

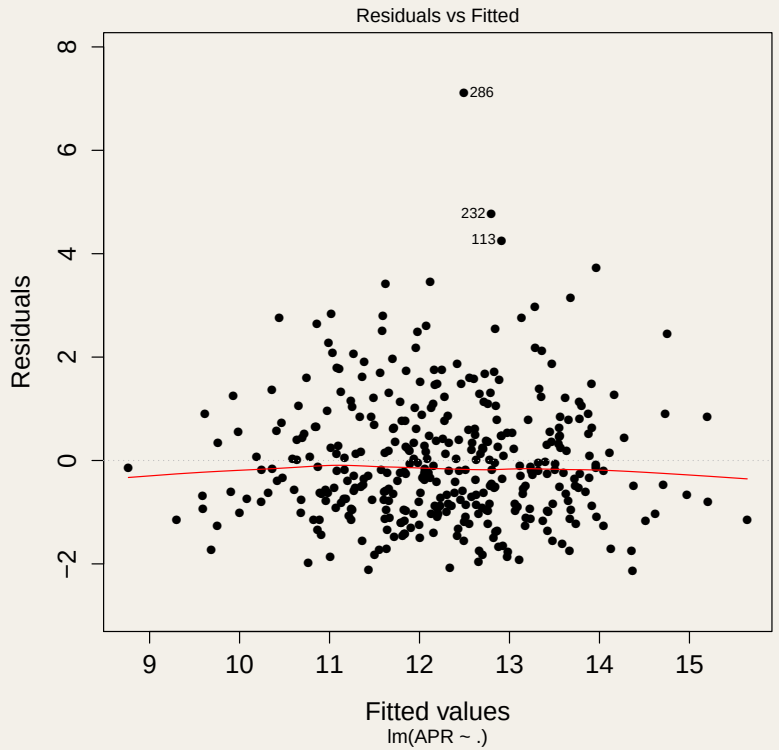


جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

4M: Subprime Mortages

Mechanics

Fit model and check conditions. Residuals versus fitted values.



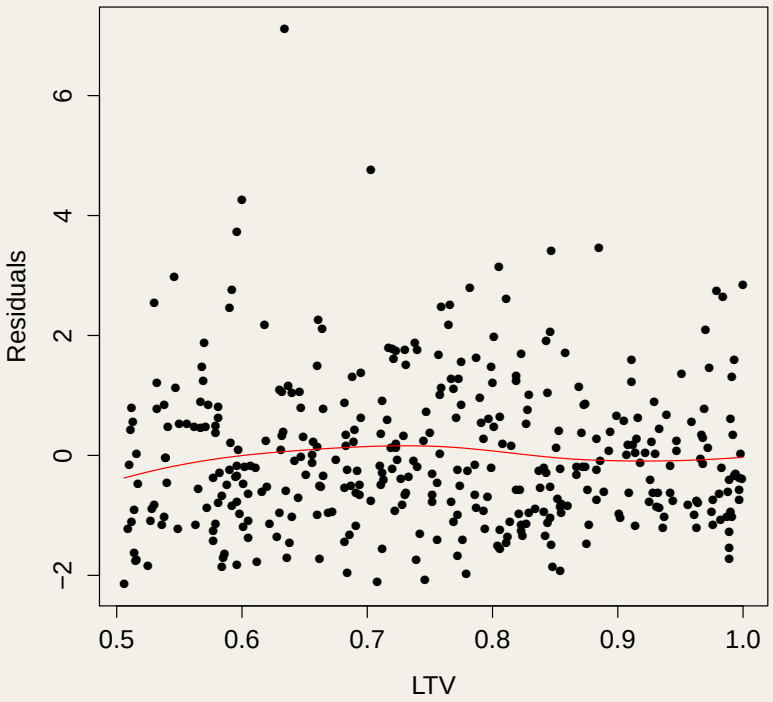
Similar variances condition is satisfied.

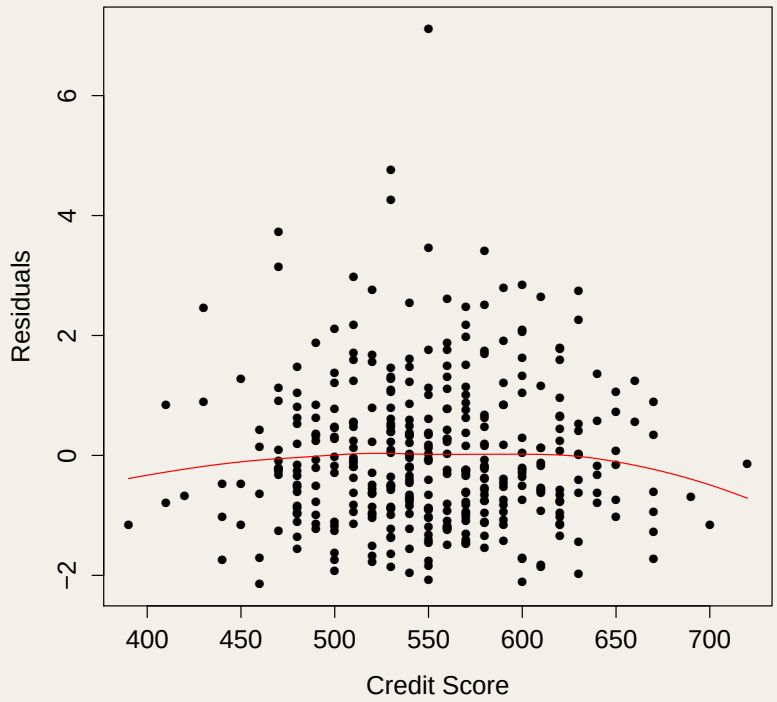
جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

4M: Subprime Mortages

Mechanics

Fit model and check conditions. Residuals versus explanatory variables.



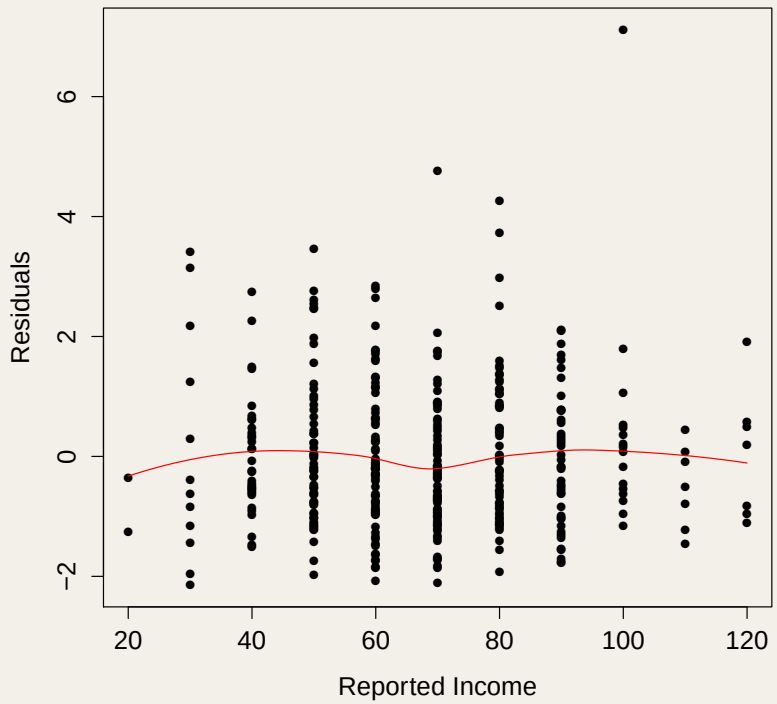


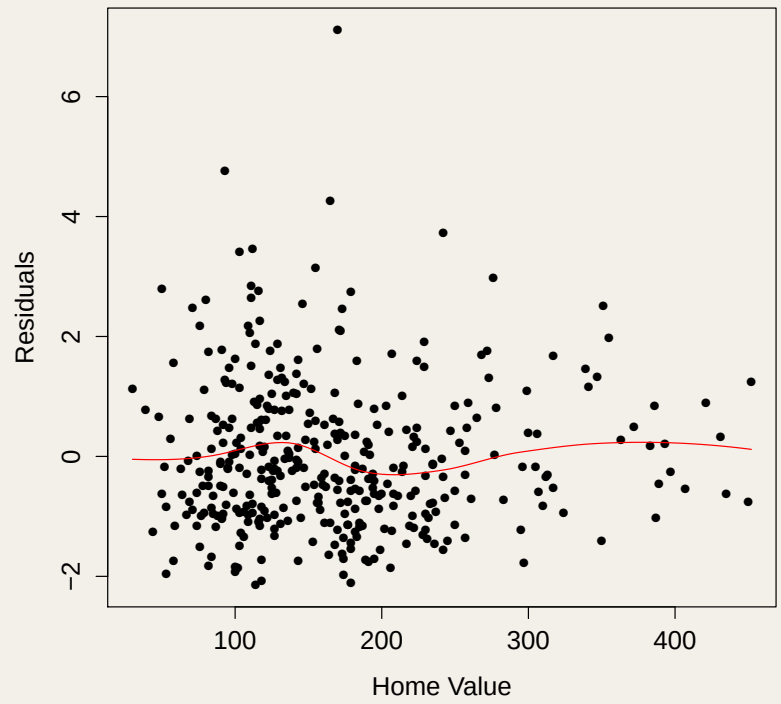
جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

4M: Subprime Mortages

Mechanics

Fit model and check conditions. Residuals versus explanatory variables.



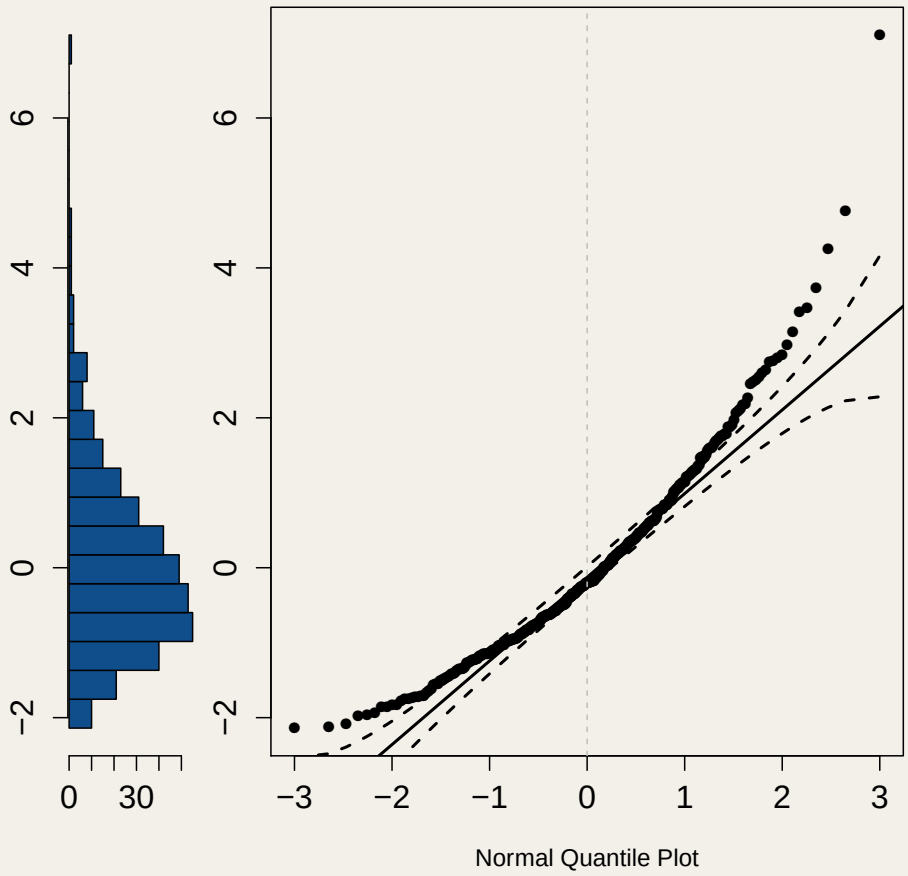


جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

4M: Subprime Mortages

Mechanics

Fit model and check conditions. Normality.

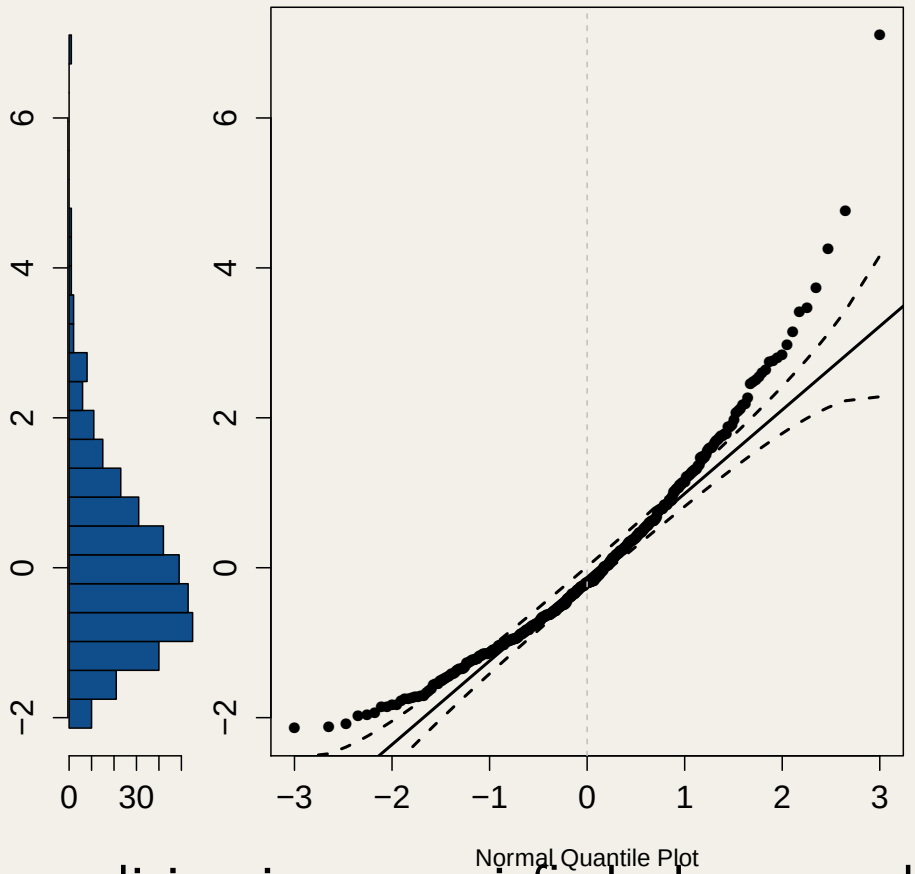


جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

4M: Subprime Mortages

Mechanics

Fit model and check conditions. Normality.



Nearly normal condition is not satisfied; data are skewed.

جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

4M: Subprime Mortages

Message

Regression Analysis shows that the characteristics of the borrower (credit score) and loan LTV affect interest rates in the market. These two factor together explain almost half the variation in interest rates. Neither income of the borrower nor the home value improves a model with these two variables.

Best Practices

- Now the context of your model
- Examine plots of the overall model and individual explanatory variables before interpreting the output.
- Check the overall F-statistic before looking at the t-statistics
- Distinguish marginal from partial slopes.
- Let your software compute prediction intervals in multiple regression.

Pitfalls

- Do not confuse multiple regression with several simple regressions.
- Do not become impatient.
- Do not believe that you have all the important variables.
- Do not think that you have found casual effects.
- Do not interpret an insignificant t-statistic to mean that an explanatory variable has no effect.
- Do not think that the order of the explanatory variables in a regression matters.