Linear Regression I

Sean Hellingman ©

Introduction to Statistical Data Analysis (ADSC1000)

shellingman@tru.ca

Fall 2024

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- We have covered correlation (ρ) as a measure of the strength of a linear relationship between two numeric variables.

- We can not use (ρ) for prediction as we do not assume a causality direction.

- We may be interested in predicting the value of a dependent variable from the value of one or more independent variables.
    - Predict the market value of a house based on the size of the home.
    - Predict students' class scores as a function of several characteristics.

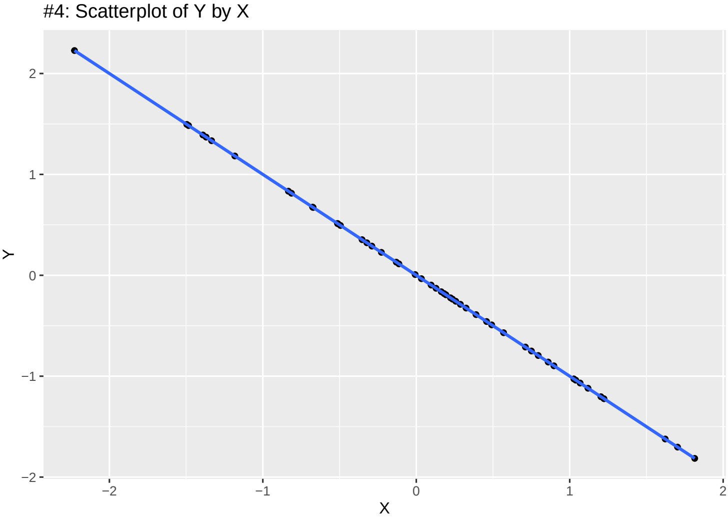
Regression Analysis

- **Regression analysis** is a statistical tool used to model relationships between a dependent variable and one or more independent (explanatory) variables.
- Right now, independent variables will be *numeric*.
- Topics covered:
  - Develop and analyse regression models with one or more continuous independent variables.
  - Basic understanding of the assumptions of regression models.
  - Interpreting results.
  - Decision-making.
  - Statistical/practical issues.

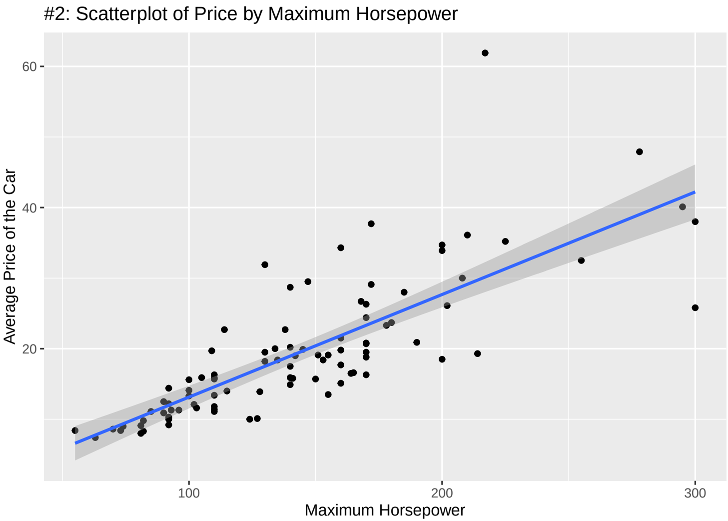
Nature of the Relationship

- One variable (Y) is the dependent (response) variable and other variables play the role of independent (explanatory) variables (X_1, X_2, ...)
- The relationship is not deterministic (functional) but is **statistical** (stochastic).
- There is a (conditional) distribution of the dependent variable associated with various combinations of independent (explanatory) variables.
- *Initially we will focus on linear relationships.*

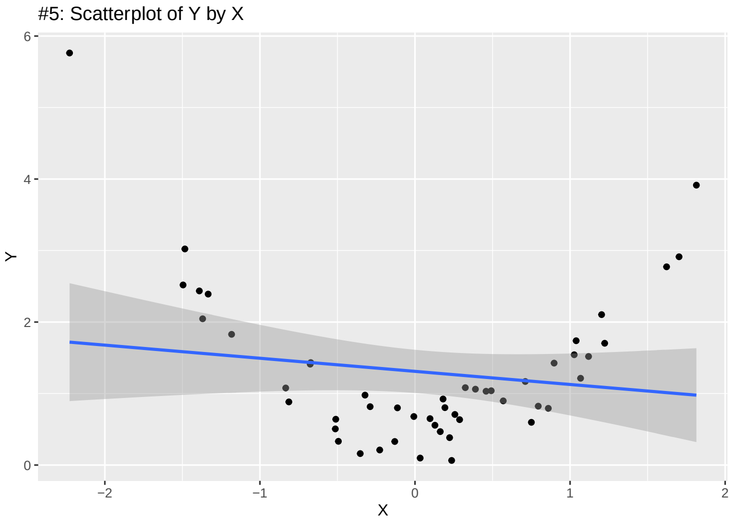
Deterministic Relationship



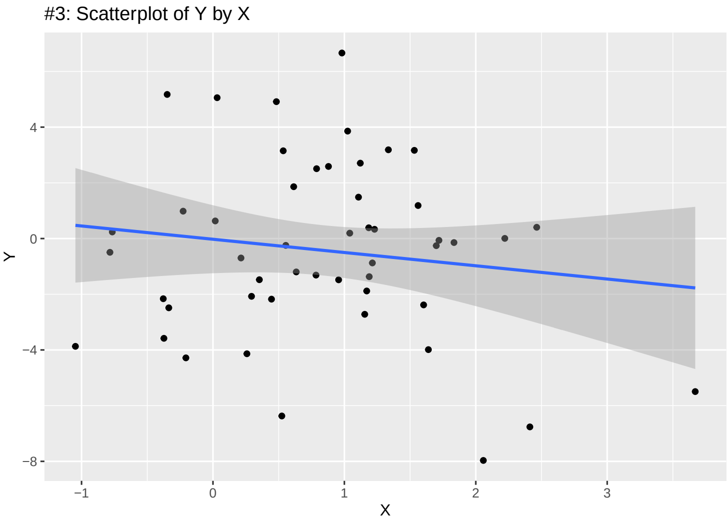
Linear Stochastic Relationship



Non-Linear Stochastic Relationship



No Stochastic Relationship



Equation of a Line

- Linear regression is based on estimating the linear relationship between the dependent and independent variable(s).
- Recall the equation of a line:

$$y = mx + b. \tag{1}$$

  - m is the slope
  - b is the y-intercept

Linear Regression Models

- Linear regression is based on estimating the linear relationship between the dependent and independent variable(s) **plus an error term**.

- **Simple linear regression model** (Expected value of Y):

$$Y = \beta_0 + X\beta_1 + \epsilon. \tag{2}$$

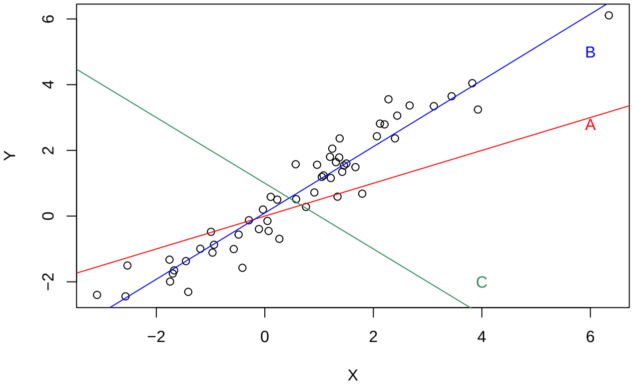
  - Y is the dependent variable.
  - β_0 is the intercept.
  - X is the independent variable.
  - β_1 is the slope of the linear relationship.
  - ϵ is the random error term.
    - Follows an assumed distribution with $E[\epsilon] = 0$ and constant variance σ_ϵ^2

Estimation

- We do not know the true values of β_0 and β_1 because we do not have the entire population.
- We need to *estimate* these parameters the best we can using the data we do have.
- If we draw a straight line, we will never be able to include all of the data points unless we have a deterministic relationship.

Example 1

- Which line should we choose to represent the linear relationship between X and Y?



Estimated Regression Line

- **The estimated simple linear regression equation is**:

$$\hat{Y} = b_0 + Xb_1. \tag{3}$$

  - b_0 and b_1 are estimates of β_0 and β_1.

- If (X_i, Y_i) is the i^{th} observation then $\hat{Y}_i = b_0 + b_1 X_i$ is the estimated value of Y for X_i.

Residuals

- One way to quantify the relationship between each point and the estimated regression equation is to measure the vertical distance between them.
- The **residuals** (observed errors) are defined as follows:

$$e_i = Y_i - \hat{Y}_i. \tag{4}$$

- The best-fitting line should minimize some measure of these errors.

Residuals Image

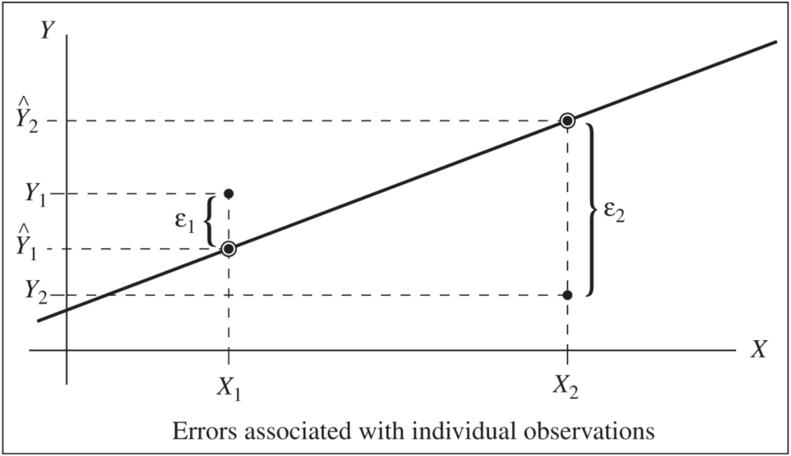


Figure: Source: (1)

Squared Residuals

- Because some of the residuals are positive and others are negative we square them (mathematical simplicity).

- We want to minimize the sum of the squared **residuals** (observed errors):

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \tag{5}$$

- The best-fitting line finds the intercept and slope that minimizes this sum (*least squares regression*).

Parameter Estimates

- Using calculus we can derive the following *least squares* estimates:

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}. \tag{6}$$

$$b_0 = \bar{Y} - b_1\bar{X}. \tag{7}$$

- R has the functionality we need to estimate these parameters.

Parameter Estimates in R

- We can estimate linear regression models in R using:
  - `lm1 <- lm(formula = y.variable ~ x.variable, data = data.frame)`
- *We can add more independent variables to the equation using +*
- To view the results of your model:
  - `summary(lm1)`

Model Summary I

```
Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.25316 -0.29087  0.03779  0.36510  1.16111 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.09559    0.07781   1.229    0.225    
X            1.00891    0.04054  24.885   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5415 on 53 degrees of freedom
Multiple R-squared:  0.9212,	Adjusted R-squared:  0.9197 
F-statistic: 619.3 on 1 and 53 DF,  p-value: < 2.2e-16
```

Model Summary II

```
Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max
-1.25316 -0.29087  0.03779  0.36510  1.16111

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09559    0.07781   1.229    0.225
X            1.00891    0.04054  24.885   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5415 on 53 degrees of freedom
Multiple R-squared:  0.9212,	Adjusted R-squared:  0.9197
F-statistic: 619.3 on 1 and 53 DF,  p-value: < 2.2e-16
```

Coefficient Estimates

Significance of coefficient estimates

Percentage of variance in Y explained by X

Model significance. Is this model better than an empty model (no explanatory variables)

Example 2

- Using the *Football22.csv* data complete estimate the following regression models:
  1. Dependent Variable (Y): *Points*
     Explanatory Variable (X): *Goals_For*
  2. Dependent Variable (Y): *Points*
     Explanatory Variable (X): *Losses*
  3. Dependent Variable (Y): *Points*
     Explanatory Variable (X): *Draws*
  4. Dependent Variable (Y): *Points*
     Explanatory Variable (X): *Goal_Differential*
- Comment on the parameter estimates and significance of your models.

Regression Assumptions and Diagnostics

Model Assumptions

- **The validity of the significance of our regression model estimates depends on some key assumptions:**

1. **Linearity**
   - The relationship between the dependent and independent variable(s) needs to be linear.
2. **Normality** (multivariate normal for multiple independent variables)
   - In linear regression, all variables must be normally distributed (can be fixed).
3. **Homoscedasticity**
   - The variation about the regression line is constant for all values of the independent variable(s) (can be fixed).
4. **Independence**
   - There is little or no multicollinearity in the data (independent variables are too highly correlated with each other).

Diagnostics

- Due to the assumptions imposed on the error term (closely related to the residuals) we can use the residuals to help us check the model assumptions.
- We can also *standardize* the residuals to help with outlier identification and assumption checks.
- **Standardized residuals**: $\text{Residual}_i/\text{Standard Deviation of Residual}_i$
- In R:
  - `rstandard(linear.model)` from the *car* package

Linearity Check

- The **linearity** assumption may be checked in a few different ways:
  1. Examine the scatterplot(s) of the dependent and independent variable(s) (linear relationship?).
  2. Plot the residuals, they should be randomly scattered around zero with no apparent pattern.
     - In R: `plot(lm1$residuals)`
     - Add line at 0: `abline(h=0,col="blue")`
  3. Plot residuals vs fitted values, again should be randomly scattered around zero with no apparent pattern.
     - In R: `plot(model, 1)`

Example 3

- Which of the linear regression models that were estimated in Example 2 pass the linearity assumption?

Normality Check

- The **normality** assumption may be checked in a few different ways (generally we focus on the residuals):
  1. Examine the histogram of the *standardized residuals* for approximate normality.
  2. Q-Q plot of the *standardized residuals*.
     - In R we can also use: `plot(model, 2)`
  3. We can use a K-S test or a Shapiro–Wilk test on the *standardized residuals*.
     - K-S: `ks.test(standardized.residuals, "pnorm")`
     - Shapiro–Wilk: `shapiro.test(standardized.residuals)`

- *Slight departures from normality may be acceptable and we can also take steps to fix departures from normality.*

Example 4

- Which of the linear regression models that were estimated in Example 2 pass the normality assumption?

Homoscedasticity Check

- The **Homoscedasticity** assumption may be checked in a few different ways (generally we focus on the residuals):
  1. Examine the scatterplot(s) of the *standardized residuals* for a constant variance.
  2. Examine the scatterplot(s) of the fitted values and the square root of the *standardized residuals* (*scale-location plot*). (It's good if you see a horizontal line with equally spread points)
     - In R: `plot(model, 3)`
  3. We can use the `ncvTest()` similar to a (Breusch–Pagan test) in R with the null hypothesis being a constant variance (Homoscedasticity).
     - In R (*car* package): `ncvTest(lm1)`
     - If we do not reject the null hypothesis (large p-value) we can assume Homoscedasticity.
- *We can also take steps to fix departures from Homoscedasticity.*

Example 5

- Which of the linear regression models that were estimated in Example 2 pass the homoscedasticity assumption?

Independence Check

- *This assumption may be violated if we have time-dependent independent (explanatory) variables.*

- The **Independence** assumption may be checked in a few different ways (generally we focus on the residuals):
  1. Examine the scatterplot(s) of the *standardized residuals* for patterns or obvious clusters.
  2. We can use the Durbin Watson test in R with the null hypothesis being that the residuals are independent.
     - In R (*car* package): `durbinWatsonTest(lm1)`
     - If we do not reject the null hypothesis (large p-value) we can assume independence at one lag.

- *There are other tests like the Ljung-Box test that may be used for multiple lags.*

Example 6

- Which of the linear regression models that were estimated in Example 2 pass the independence assumption?

Thoughts

- When any of the assumptions are violated, our inferences may not be valid.
    - **It is very important to check the validity of the assumptions.**
- There are steps we can take to improve things if one or more of the assumptions is violated.
- We can use `plot(lm1)` in R to examine some of the plots we covered all at once.
- We should also be cautious of outlier influences on our model estimates.

Exercise 1

- Load the *Cars93* dataset from the *MASS* R package and create simple linear regression models to evaluate the relationships between the following pairs of dependent and independent variables:

  1. *Price* and *RPM*
  2. *Price* and *EngineSize*
  3. *Horsepower* and *EngineSize*
  4. *Weight* and *Fuel.tank.capacity*
  5. *Weight* and *Width*

- Comment on the estimated slopes and intercepts.

- Be sure to check the validity of the four assumptions we covered.

Exercise 2

- Take some time to create and validate some simple linear regression models using your project data.
- **Keep in mind that causality should be one-directional!**
    - Changes in the dependent variable should be caused by changes in the independent (explanatory) variable and not the other way around.

References & Resources

1. Evans, J. R., Olson, D. L., & Olson, D. L. (2007). *Statistics, data analysis, and decision modeling.* Upper Saddle River, NJ: Pearson/Prentice Hall.

2. Devore, J. L., Berk, K. N., & Carlton, M. A. (2012). *Modern mathematical statistics with applications (Second Edition).* New York: Springer.

- https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm
- https://www.rdocumentation.org/packages/car/versions/1.2-6/topics/ncv.test
- https://www.rdocumentation.org/packages/car/versions/3.1-2/topics/durbinWatsonTest
- https://cran.r-project.org/web/packages/car/index.html