Social Robotics

Albert-Ludwigs-Universität Freiburg

Bernhard Nebel, Felix Lindner, Thorsten Engesser,
Barbara Kuhnert, Laura Wächter

WS 2017/18

Linear Regression

Basic Setting

- Linear regression allows to study relationships between independent and dependent variables.
- We will focus on simple linear regression, which models the relationship between one independent and one dependent variable.
  - The independent variable is called the predictor.
  - The dependent variable is called the response.

The Running Example

- Imagine we want to model the relationship between a social robot's politeness and its perceived likability. We assume politeness to be configurable by the robot's administrator. To study the relationship, a data sample has been gathered, which looks like this:
  - politeness <- c(10, 10, 20, 20, 30, 30, 40, 40, 50, 50, 60, 60)
  - likability <- c(3, 8, 9, 10, 7, 9, 9, 11, 11, 12, 10, 13)
- Questions we might want to answer:
  - What is the mean politeness of the robot?
  - What is the likeability score of the robot with politeness 45?

The Mean as Simplest Model

- The Model
  - Model: $Y_i = \overline{Y} + \varepsilon_i$
  - Residuals: $\varepsilon_i = Y_i - \overline{Y}$
  - Sum of Squared Error: $SSE = \sum_i^n \varepsilon_i^2$
  - Mean Squared Error: $MSE = SSE/(n-1)$
  - Residual Standard Error: $S = \sqrt{MSE}$ (this is just s^2)
- Running example
  - $\overline{Y} = 9.33$. The model will claim that the average robot's perceived likability is 9.33 regardless of its politeness. Similarly, if we use this model to answer the question about the likability of the robot with politeness 45, it will answer 9.33.

⇒See `lecture13.Rmd` for an example

The Simple Linear Model I

- The Model
  - Model: $Y_i = b_0 + b_1 X_i + \varepsilon_i$, with Intercept: b_0 and Slope: b_1.
  - Residuals: $\varepsilon_i = Y_i - b_0 - b_1 X_i$
  - Sum of Squared Error: $SSE = \sum_i^n \varepsilon_i^2$
  - Mean Squared Error: $MSE = SSE/(n-2)$
    - Two variables get estimated (the intercept and the slope), thus two degrees of freedom.
  - Residual Standard Error: $S = \sqrt{MSE}$
- Running example
  - $Y_i = 5.533 + 0.11 X_i$. The model will claim that the average robot's perceived likability with politeness 45 will be $5.533 + 0.11 \cdot 45 = 10.42$. Similarly, if we use this model to answer the question about the likability of the robot with politeness 45, it will answer 10.42.

⇒See `lecture13.Rmd` for an example

The Simple Linear Model II

- In a nutshell, given a linear model $Y_i = b_0 + b_1 X_i + \varepsilon_i$:
- $\widehat{Y}_i = b_0 + b_1 X_i$ represents the mean response at position X_i,
- the error term ε_i is assumed to be normally distributed with mean 0 and equal variances for all levels of X.

Least Squares Fitting I

- Goal: Determination of the best fitting line
- Notation
  - Y_i: ith observed response
  - X_i: ith predictor value
  - $\widehat{Y}_i$: ith predicted response (aka the fitted value)
- The equation for the best fitting line is:

$$\widehat{Y}_i = b_0 + b_1 X_i$$

- Coefficients b_0 and b_1 are to be found such that they minimize:

$$Q = \sum_i^n (Y_i - \widehat{Y}_i)^2$$

Least Squares Fitting II

- Coefficients b_0 and b_1 are to be found such that they minimize:

$$Q = \sum_i^n (Y_i - \widehat{Y}_i)^2 = \sum_i^n (Y_i - (b_0 + b_1 X_i))^2$$

- Derivations of Q with respect to b_0 and b_1 are taken, set to 0, and solved for b_0, b_1, resulting in:

$$b_0 = \overline{Y} - b_1 \overline{X}$$

$$b_1 = \frac{\sum_i^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i^n (X_i - \overline{X})^2}$$

Intercept and Slope

- Intercept: $b_0 = \overline{Y} - b_1 \overline{X}$
- Slope: $b_1 = \frac{\sum_i^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i^n (X_i - \overline{X})^2}$
- Compare the coefficient b_1 to Pearson's r: $r = \frac{Cov(X,Y)}{s_X s_Y}$
  - $b_1 = \frac{Cov(X,Y)}{s_X s_X} = r \frac{s_Y}{s_X}$
  - Thus, if X, Y are standardized variables (mean zero, standard deviation 1), then: $b_1 = r$, $b_0 = 0$, $\widehat{Y}_i = rX_i$.

On b_0 and b_1

- b_0: This coefficient is very often meaningless, especially when 0 is out of X's range. See `lecture13.Rmd` for an example.
- b_1: It's the increase of the mean response for every one unit increase in X.

Coefficient of Determination R^2

Basic Terminology

- Regression sum of squares: $SSR = \sum_i^n (\widehat{Y}_i - \overline{Y})^2$
  - Quantifies how far the estimated slope regression line, $\widehat{Y}$, is from the horizontal "no relationship" line, $\overline{Y}$.
- Sum of squared error: $SSE = \sum_i^n (Y_i - \widehat{Y}_i)^2$
  - Quantifies how much the data points Y_i vary around the estimates regression line $\widehat{Y}$.
- Total sum of squares: $SSTO = SSR + SSE = \sum_i^n (Y_i - \overline{Y})^2$
  - Quantifies how much the data points Y_i vary around their mean $\overline{Y}$.
- Thus, one can say that the regression line absorbs some of the variance (i.e., SSR) from the total variance $SSTO$ and leaves SSE unexplained.

Coefficient of Determination R^2

Properties

- R^2 value is the regression sum of squares divided by the total sum of squares:

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

- R^2 is a number between 0 and 1.
- If $R^2 = 1$, i.e., $SSE = 0$: The predictor X explains all of the variation in Y .
- If $R^2 = 0$, i.e., $SSR = 0$, i.e., $\widehat{Y} = \overline{Y}$: The predictor X explains none of the variation in Y.
- $R^2 \cdot 100$ percent of the variation in Y is explained by X.

Intermediate Summary

- $Y_i = b_0 + b_1 X_i$ is the model for the mean response Y_i on a given X_i.
- Coefficients b_0 and b_1 can be computed using the equations from slide 9.
- R^2 can be used to assess how much of the total variation in the data can be explained by the predictor.
- Next: How can we test that the contribution of the predictor is statistically significant? That is, is b_1 statistically different from 0?

The t Statistics for the Slope

- Test for hypothesis $H1 : \beta_1 \neq 0$, $H0 : \beta_1 = 0$
- If there is a linear relationship in the population with slope β_1, then the b_1s are normally distributed with mean β_1. The variance can be estimated by $\sigma^2 = \frac{S}{\sqrt{\sum_i^n (X_i - \overline{X})^2}}$ with residual standard error S.

$$t = \frac{b_1 - \beta}{\frac{S}{\sqrt{\sum_i^n (X_i - \overline{X})^2}}} = \frac{b_1 - 0}{\frac{S}{\sqrt{\sum_i^n (X_i - \overline{X})^2}}} = \frac{b_1}{\frac{\sqrt{\frac{\sum_i^n (Y_i - \widehat{Y})^2}{(n-2)}}}{\sqrt{\sum_i^n (X_i - \overline{X})^2}}} = \frac{b_1}{s_{b_1}} \sim t(n-2)$$

⇒See slide `lecture13.Rmd` for an example.

Relation to Two-Sample t-Test

Motivation

- In the beginning we said that the $\widehat{Y}_i$ are the mean response for each X_i, and that b_1 represents the increase of this response if X is increased by one unit.
- So, what if X is just binary (0, 1)?
- Then, $\widehat{Y}_i = b_0$ is the mean response for $X = 0$, and $\widehat{Y}_i = b_0 + b_1$ is the mean response of $X = 1$.
- Using a t-Test on b_1, we can test the significance of this increase.
- Isn't this just a Two-Sample t-Test?

Relation to Two-Sample t-Test

Derivation I

- $b_1 = \frac{\sum_i^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i^n (X_i - \overline{X})^2}$, $b_0 = \overline{Y} - b_1 \overline{X}$
- We encode the two groups with a binary variable X with values 0 and 1. Hence, $\overline{X} = \frac{1}{2}$. We further assume that both the groups have equal number of samples, thus $\overline{Y} = \frac{\overline{Y_0} + \overline{Y_1}}{2}$.
- $b_1 = \frac{\sum_i^n (X_i - \frac{1}{2})(Y_i - \overline{Y})}{\frac{n}{4}} = \frac{\frac{1}{2}\sum_i^{n/2}(Y_{1,i} - \overline{Y}) - \frac{1}{2}\sum_i^{n/2}(Y_{0,i} - \overline{Y})}{\frac{n}{4}} = 2\frac{\sum_i^{n/2}(Y_{1,i} - \overline{Y}) - \sum_i^{n/2}(Y_{0,i} - \overline{Y})}{n} = \frac{\sum_i^{n/2}(2Y_{1,i} - \overline{Y_0} - \overline{Y_1}) - \sum_i^{n/2}(2Y_{0,i} - \overline{Y_0} - \overline{Y_1})}{n} = \frac{\sum_i^{n/2}(Y_{1,i} - \overline{Y_0}) - \sum_i^{n/2}(Y_{0,i} - \overline{Y_1})}{n} = \frac{\overline{Y_1} - \overline{Y_0}}{2} + \frac{\sum_i^{n/2}(Y_{1,i} - Y_{0,i})}{n} = 2\frac{\overline{Y_1} - \overline{Y_0}}{2} = \overline{Y_1} - \overline{Y_0}$
- $b_0 = \frac{\overline{Y_0} + \overline{Y_1}}{2} - \frac{b_1}{2} = \overline{Y_0}$
- Hence, $\widehat{Y}_i = \overline{Y_0} + (\overline{Y_1} - \overline{Y_0})X_i$

Relation to Two-Sample t-Test

Derivation II

- Reconsider $t = \frac{b_1}{\frac{\sqrt{\frac{\sum_i^n (Y_i - \widehat{Y})^2}{(n-2)}}}{\sqrt{\sum_i^n (X_i - \overline{X})^2}}} \sim t(n-2)$
- $t = \sqrt{\frac{n}{2}} \frac{b_1}{\sqrt{2\frac{\sum_i^n (Y_i - \widehat{Y})^2}{(n-2)}}} = \sqrt{\frac{n}{2}} \frac{\overline{Y_1} - \overline{Y_0}}{\sqrt{2\frac{\sum_i^n (Y_i - \widehat{Y})^2}{(n-2)}}} = \sqrt{\frac{n}{2}} \frac{\overline{Y_1} - \overline{Y_0}}{\sqrt{2\frac{(n-1)s^2_{Y_0} + (n-1)s^2_{Y_1}}{(n-2)}}} = \sqrt{\frac{n}{2}} \frac{\overline{Y_1} - \overline{Y_0}}{\sqrt{s^2_{Y_0} + s^2_{Y_1}}} \sim t(n-2)$
- This is exactly the term from our t-Test lecture! (The only syntactical difference is that we wrote $N = n/2$)
- In case of a binary predictor, running a Two-Sample Paired t-Test is just the same as running a t-Test on the slope of a linear regression model.

The F Statistics for the Model I

- The R summary of a linear regression model also outputs the F statistics along with a p value. Intuitively, this test assesses if the regression model makes a significant contribution to explain the data (rather than testing individual contributions of the coefficients).
- However, in case of only one predictor, this is just the same as testing $H1 : \beta_1 \neq 0, H0 : \beta_1 = 0$. And therefore, we expect the same result as with the t-Test on the same hypothesis.

The F Statistics for the Model II

- Remember the ANOVA for testing differences between several means: The idea was to compare the variance of the means to the total variance.
- For regression, something similar is done: The Regression sum of squares (SSR) is compared to the mean Residual error (SSE/(n-2)).
  - Regression sum of squares: $SSR = \sum_i^n (\widehat{Y}_i - \overline{Y})^2$
  - Sum of squared error: $SSE = \sum_i^n (Y_i - \widehat{Y}_i)^2$
  - Total sum of squares: $SSTO = SSR + SSE = \sum_i^n (Y_i - \overline{Y})^2$
- $F = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}}$, with $df1 = 1, df2 = n - 2$.
- F becomes bigger as the regression line becomes steeper.
- Hence, if the groups are appropriately modeled, the F test for $H0 : \beta_1 = 0$ is equal to the F test for $H0 : \mu_0 = ... = \mu_n$, for which we used the ANOVA. In fact, the aov() procedure in R just calls lm().

Report

To report the result from a simple linear regression, you can write:

> *A simple linear regression was calculated to predict [dependent variable] based on [independent variable]. A significant regression equation was found (F([df1], [df2]) = [F value], p = [p value]), with an R^2 of [R-Squared value].*

You may add:

> *[Dependent variable] is equal to [intercept] + [b1] (independent variable) [dependent variable measure] when [independent variable] is measured in [unit of measure].*

Correlation vs. Regression

When to report what?

- If one of the variables can be clearly identified as the response of another variable, then run a linear regression and report the t-test or F-test results for $H0 : \beta_1 = 0$.
  - In case of the simple linear regression, both these tests are equivalent.
  - For multi-regression, the F-test tests if one slope is significant, and the t-tests test the significance of the particular slopes.
- If it is not obvious which of the variables is the response (i.e., you do not want to assume causality), then run a correlation and report a t-test for $H0 : r = 0$.

This closes the mathy part of the lecture.

Sketches

Intentionally left blank :-)