McGill University

MATH 204

Class Notes

Alexandre St-Aubin

Taught by
Jose Andres Correa

September 3, 2023

Contents

1 Simple linear regression (lectures 1-2)

Definition 1.1. (Simple linear regression) Let Y_i be the response variable which depends on X_i (**Explanatory variable**), for $i \in \{1, 2, ..., n\}$.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1}$$

where β_0 is the y-intercept, β_1 is the slope, and ϵ_i is the model (random) error.

Note 1.1. β_0 and β_1 are population parameters, their values need to be estimated. We will denote the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. The values of the regression line are called predicted values;

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{2}$$

Definition 1.2. (Least squares criterion) This criterion minimizes the distance between each $\hat{y}_i$ and y_i.

$$\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i)^2 \tag{3}$$

where

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \tag{4}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{5}$$

Where $\bar{y}$ and $\bar{x}$ represent the mean of the observed data. These values of $\hat{\beta}_0$ and $\hat{\beta}_1$ give the minimum value of the squared differences between the observed and fitted values. S_{XY} corresponds to the sample covariance, S_{XX} the standard deviation of x_i, and S_{YY} the standard deviation of y_i.

Note 1.2. The error terms $\epsilon_1, ..., \epsilon_n$ in equation (1) are assumed to be mutually independent random variables with mean 0 and variance σ^2.

Note 1.3. $Y_1, ..., Y_n$ are random variables, mutually independent, and for each $i \in \{1, ..., n\}$,

$$E(Y_i | X_i = x_i) = \beta_0 + \beta_1 x_i \tag{6}$$

$$Var(Y_i | X_i = x_i) = \sigma^2 \tag{7}$$

So, β_1 corresponds to the change in the mean of Y_i for a one unit change in x_i. Thus if you compare two experimental units whose x_i variables differ by 1, the difference in the means of those two experimental units is β_1. On the other hand, β_0 is the mean of Y_i when $x_i = 0$.

Definition 1.3. (Unbiased estimator) We say $\hat{\beta}_1$ is an unbiased estimator of β_1 if

$$E(\hat{\beta}_1) = \beta_1 \tag{8}$$

That is, if used over and over again on a large number of different data sets, the formula for $\hat{\beta}$ will tend to give a faithful estimate of the unknown population value β on average.

Definition 1.4. (Standard deviation of β_1)

$$\sigma_{\hat{\beta}_1} = \sqrt{Var(\hat{\beta}_1)} = \frac{\sigma}{S_{XX}} \tag{9}$$

From which we can derive

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{XX}} \tag{10}$$

We can estimate σ^2 (variance of residuals) with

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{SSE}{n-2} \tag{11}$$

$(n-2)$ is the error or residual degrees of freedom.

$$SSE = S_{YY} - \hat{\beta}_1 S_{XY} \tag{12}$$

Estimating the standard deviation

$$\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}} = \frac{\hat{\sigma}}{\sqrt{S_{XX}}} \tag{13}$$

2 Confidence interval for β_1 (lecture 3)

Proposition 2.1. Suppose the error terms $(\epsilon_1, ..., \epsilon_n)$ are normally distributed,

$$(\epsilon_1, ..., \epsilon_n) \sim \mathcal{N}(0, \sigma^2) \tag{14}$$

Then, $\hat{\beta}_1$ is normally distributed,

$$\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \sigma^2/S_{XX}) \tag{15}$$

Definition 2.1. (Random variabe T)

$$T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{XX}}} \sim t_{n-2} \tag{16}$$

where t_{n-2} is the Student random distribution with $n-2$ degrees of freedom.

Definition 2.2. (Confidence Interval) To form a $100 \times (1-\alpha)\%$ confidence interval for β_1, one can use

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} \times \hat{\sigma}_{\hat{\beta}_1} = \hat{\beta}_1 \pm t_{n-2,\alpha/2} \times \frac{\hat{\sigma}}{\sqrt{S_{XX}}} \tag{17}$$

Example 2.1. Therefore, for Type I error rate $\alpha = 0.05$ we would reject H_0 for values of the test-statistics larger than 1.96 or smaller than -1.96, i.e., $RR = |T| > 1.96$.

Note 2.1. A type I error is said to result if the test rejects H_0 when H_0 is true. A type II error is said to occur if the test does not reject H_0 when H_0 is false.

3 P-value (lecture 4)

Definition 3.1. (p-value) The p-value tells you how likely it is that your data could have occurred under the null hypothesis. The p value tells you how often you would expect to see a test statistic as extreme or more extreme than the one calculated by your statistical test if the null hypothesis of that test was true. The p value gets smaller as the test statistic calculated from your data gets further away from the range of test statistics predicted by the null hypothesis. Here are some general guidelines:

1. $p < 0.001$; extremely strong evidence
2. $0.001 \leq p < 0.01$; very strong evidence
3. $0.01 \leq p < 0.05$; strong evidence
4. $0.05 \leq p < 0.10$; modest evidence
5. $p > 0.10$; poor evidence

Example 3.1. The p value is a proportion: if your p value is 0.05, that means that 5% of the time you would see a test statistic at least as extreme as the one you found if the null hypothesis was true. **In linear regressions, the null hypothesis is that $\beta_1 = 0$.**

Theorem 3.1.

$$T \sim t_v \implies T^2 \sim F(1, v)$$

where $F(\alpha, \beta)$ refers to the Fisher-Snedecor distribution.

4 Correlation and Coefficient of Determination (lecture 5)

Definition 4.1. (Correlation) The correlation between two random variables, X and Y (or the correlation coefficient) can be calculated as

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} \tag{18}$$

The correlation coefficient, r, lies between -1 and $+1$ and it is scaleless. The correlation coefficient r is an estimator for the population p.

Definition 4.2. (Fisher's variance stabilizing z-transformation) This method can be used to build a confidence interval for p. First, transform r to $z = \frac{1}{2}ln(\frac{1+r}{1-r})$. Then, build a confidence interval:

$$z \pm \frac{z_{\alpha/2}}{\sqrt{n-3}} = (c_l, c_u) \tag{19}$$

where $z_{\alpha/2}$ is from the standard Normal table. Now reverse the transformation and report a $100 \times (1-\alpha)\%$ confidence interval for p of

$$\left(\frac{e^{2c_L} - 1}{e^{2c_L} + 1}, \frac{e^{2c_U} - 1}{e^{2c_U} + 1} \right) \tag{20}$$

Definition 4.3. (Coefficient of Determination) The coefficient of determination, denoted by R^2, is defined by:

$$R^2 = 1 - SSE/S_{YY} \tag{21}$$

R^2 gives the proportion of variance of Y explained by X.

Note 4.1. If X is not linearly associated with Y, $SSE = S_{YY} \implies R^2 = 0$. If X is linearly associated with Y, $SSE < S_{YY} \implies 0 < R^2 < 1$.

Note 4.2. The square of the correlation coefficient r^2 equals the coefficient of determination R^2.

Example 4.1. Suppose you have a R^2 value of x%, then we say that x% of the variance in y is explained by the change in x.

5 Estimating y-values (lecture 6)

5.1 Estimating the mean response

Given a particular value x_0, it is possible to estimate the mean response value $\hat{y}_0$:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

where $\hat{y}_0$ is unbiased, i.e.

$$E(\hat{y}_0) = \beta_0 + \beta_1 x_0$$

and,

$$var(\hat{y}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)$$

which we can estimate with

$$v\hat{a}r(\hat{y}_0) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)$$

Definition 5.1. (Confidence interval for the mean value y_0) If the error terms are normally distributed, a $100 \times (1 - \alpha)\%$ confidence interval for the mean value y_0 is given by

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \times \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}}$$

5.2 Predicting a specific value Y_0

It is possible to estimate some individual value Y_0 for x_0.

Definition 5.2. (Variability of Y_0)

$$v\hat{a}r(Y_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)$$

Definition 5.3. (Confidence interval for the individual value Y_0) If the error terms are normally distributed, a $100 \times (1 - \alpha)\%$ confidence interval for an individual value Y_0 is given by

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \times \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}}$$

Definition 5.4. (lm summary command on R)

1. Estimate:
   - (intercept) $\hat{\beta}_0$
   - (slope) $\hat{\beta}_1$
2. Residual standard error: $\hat{\sigma}$
3. Multiple R-squared: R^2
4. p-value: trivial
5. F-statistic: $T^2{}_{obs}$
6. t value:
   - (intercept)
   - (slope) $T_{obs} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$

6 Residual Analysis (lectures 7-8)

Previous assumptions:
$\epsilon_1, ..., \epsilon_n$ are mutually independent, and for each $i \in \{1, ..., n\}$,

$$E(\epsilon_i) = 0$$

$$var(\epsilon_i = \sigma^2)$$

$$\epsilon_i \sim N(0, \sigma^2)$$

6.1 Validating the normality assumption

Definition 6.1. (Residual estimates) Recall that the fitted values for the model are

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

This is the best estimate of the mean for Y_i. Therefore, the best guess as to the unobservable value of $\epsilon_i$is

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

The quantity ϵ_i is referred to as the residual or the ith observation.

Definition 6.2. (Standardized residual estimates) It turns out that the mean of the sample residuals is always zero, so the ith standardized residuals is equal to

$$\hat{\epsilon}_i{}^{std} = \frac{\hat{\epsilon}_i}{\hat{\sigma}}$$

where $\hat{\sigma}$ is the estimate of σ from the regression.

Definition 6.3. (Empirical Rule) For Normally distributed observations, about 95% of the observations lie within 2 standard deviations from the mean, and almost all observations lie within 3 standard deviations from the mean.

Definition 6.4. (Regression Outliers) Studentized residuals have mean 0 and standard deviation 1. Therefore under the assumption that they are Normally distributed, residuals that are bigger than 3 in absolute value are possible **regression outliers**. In the case of a regression outlier, if the data point is not "special" or "different," then it should not be excluded.
One solution: analyze the data both with and without the outlier to see if the conclusions change much...

Technique
Either use a Q-Q plot, or a histogram to check wether the standardized residual estimates are within the standard deviation.

6.2 Validating the variance and estimated value

Technique: to check these two assumptions, one can plot the residuals (standardized or not) against the fitted values. Then,

1. If the residuals are truly all mean zero, then one should not see any residuals that vary as a function of the fitted mean.

2. If the residuals all have the same variance, one should see equal variability across all the fitted values, as opposed to heteroscedasticity .

7 Polynomial regression (lecture 9)

Definition 7.1. (**Polynomial Regression**) In practice, it is obvious that not all data follows a simple linear regression model. This is where the polynomial regression model comes in handy. In general, polynomial regression specifies that

$$E(Y|X = x) = \beta_0 + \beta_1 + \ldots + \beta_p x^p$$

where the integer p is the largest power of x in the model. All intermediate powers need not be present.

7.1 Estimating the parameters

The parameters $\beta_0, \ldots, \beta_p$ can be estimated by the method of least squares, i.e. by choosing $\beta_, \ldots, \beta_p$ so that

$$\sum_{i=1}^{n} = (y_i - \beta_0 - \beta_1 x_i - \ldots - \beta_p x_i^p)$$

is minimized. The estimates $\hat{\beta}_0, \ldots, \hat{\beta}_p$ can be computed easily using R. Using these estimates, we can compute the predicted or fitted values

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \ldots + \hat{\beta}_p x_i^p$$

And the residuals

$$\hat{\epsilon}_i = y_i - \hat{y}_i \ \forall i \in \{1, ..., n\}$$

Note 7.1. The diagnostic tools of residual analysis discussed earlier continue to apply verbatim.

7.2 Fitting a polynomial regression model in R

The command lm in R can be used to fit a polynomial regression model. Higher-order terms such as x^2 or x^3 are specified using the I($\cdot$) function.

$$\rightarrow sal.quad.mod = lm(Salary \sim Experience + I(Experience\hat{}2))$$

$$summary(sal.quad.mod)$$

Note 7.2. (An artifact of using the quadratic model) Any parabola $\beta_0+\beta_1 x+\beta_2 x^2$ will eventually decrease if $\beta_2 < 0$.We could try to improve the fit by adding more terms to the model, viz.

$$x^3, x^4, ...$$

Note 7.3. In fitting a polynomial regression model,

1. It is important to be sensitive both to overfitting local features of the data and also to interpretation;
2. The more complex the model is, the less plausibly can we justify trying to estimate it with a relatively small amount of data;
3. Generally speaking, it is also preferable not to have a quadratic term in the model without a linear term.

8 Multiple Variable Regression

We will continue to assume that our covariates have a linear association with the response variable. Assume that there are K covariates of interest, $X_1, ..., X_K$, and a single response variable, Y. The regression model assumes that for each $i \in 1, ..., n$, one has, conditional on covariate values $x_{i1}, ..., x_{iK}$,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_K x_{iK} + \epsilon_i,$$

i.e., the response variable can be written as a non-random linear function of the covariates plus some random error.

8.1 What is linear?

The model is linear in the parameters, not necessarily in the covariates, that is linear regression (multiple or simple) refers to linearity in the coefficients of the model (β_i's), not necessarily in the covariates (x_i's). For example,

$$ln(Y_i) = \beta_0 + \beta_1 e^{x_{i1}} + \beta_2 x_{i2}^2$$

is a linear model.

8.2 Assumptions behind the multiple linear model

1. $E(\epsilon_1) = \ldots = E(\epsilon_n) = 0$
2. $var(\epsilon_1) = \ldots = var(\epsilon_n) = \sigma^2$
3. $(\epsilon_1, \ldots, \epsilon_n)$ are mutually independent.

8.3 Interpretation

Interpreting the coefficients of a multiple regression model is more complicated than in a simple linear regression model. β_j is the change in the mean of Y_i observed for a single unit increase in x_{ij}, holding all other variables $x_{i1}, ..., x_{iK}$ constant.

β_j in a regression model is not just the association of X_j with Y , but is actually the association of X_j with Y while adjusting for the associations of all the other covariates with Y .

8.4 Inference

As in simple linear regression, $\hat{\beta}$ is unbiased for β, i.e.

$$E(\hat{\beta}) = \beta$$

So, for all j,

$$E(\hat{\beta}_j) = \beta_j$$

The variance of any $\hat{\beta}_j$, $\hat{\sigma}^2_{\hat{\beta}_j}$ is difficult to compute, as the formula depends on the variance of the model errors σ^2. To get an unbiased estimate of σ^2:

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

Where $K + 1$ is the number of coefficients that are estimated in the model, $\beta_0, \ldots, \beta_K$.

8.5 Hypothesis tests

8.5.1 Testing an individual coefficient (lecture 11)

The hypotheses we are interested in are

$$\mathcal{H}_0 : \beta_j = 0,\ \mathcal{H}_a : \beta_j \neq 0$$

Remark. Failing to reject H_0 does not mean that X_j is not associated with Y. It means that it is not associated with Y after adjusting for the associations of all other variables with Y.

Testing the regression coefficients

To form a $100 \times (1-\alpha)\%$ confidence interval for β_j after adjusting for all other covariates, we do

$$\hat{\beta}_j \pm t_{n-(K+1),\,\alpha/2} \times \hat{\sigma}_{\hat{\beta}_j}$$

Remark. If 0 is in the interval it does not mean that X_j is not associated with Y . It means that it is not associated with Y after adjusting for the associations of all other variables with Y .

8.5.2 Coefficient of Determination

Definition 8.1. (Multiple coefficient of determination) R^2 is the percentage of variation in Y which is explained by the regression model.

$$R^2 = 1 - \frac{SSE}{S_{YY}}$$

R^2 is a summary of how strongly the response variable is linearly associated to the linear combination of the covariates.

Remark. Because the coefficient of determination, R^2, always increases when a new independent variable is added to the model, it is tempting to include many variables in the model in order to force R^2 to be near 1. However, doing so reduces the number of degrees of freedom available for estimating σ^2, which adversely affects our ability to make reliable inferences.

Definition 8.2. (Adjusted coefficient of determination R_a^2) Cannot be "forced" to be 1 if we include many many X's in the model.

$$R_a^2 = 1 - \frac{n-1}{n-(K+1)}\left(\frac{SSE}{S_{YY}}\right) = \frac{n-1}{n-K-1}R^2 - \frac{K}{n-K-1}$$

8.5.3 Testing the overall/global fit of the model (lecture 12)

An overall hypothesis test for the regression model. We can write the null hypothesis as

$$\mathcal{H}_0 \ : \ \beta_1 = \ldots = \beta_k = 0$$

The alternative hypothesis is that at least one β_j for the covariates is not equal to zero.

$$\mathcal{H}_a : \text{At least one } \beta_j \text{is not equal to 0}$$

The method is to use the F-statistic.

Definition 8.3. The F-statistic, under $\mathcal{H}_0$ will have an F distribution with K and $n-(K+1)$ degrees of freedom.

$$F_{obs} = \frac{R^2/K}{(1-R^2)/(n-K-1)} = \frac{MSR}{MSE}$$

We reject $\mathcal{H}_0$ for large values of F, that is when

$$F_{obs} > \mathcal{F}_{\alpha,K,n-K-1}$$

Remark. Rejecting $\mathcal{H}_0$ does not say which of coefficients are unlikely to be equal to zero, only that there is significant evidence against the hypothesis that they all are zero. It is possible to reject $\mathcal{H}_0 \ : \ \beta_1 = \ldots = \beta_k = 0$ without rejecting a single null hypothesis of the form $\mathcal{H}_0 \ : \ \beta_j = 0$.

8.6 Prediction

Suppose you have a set of values for the covariates $x_0 = (x_1, ..., x_k)$, then the value of Y predicted by the model is

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$$

The formula could be used to estimate

- The mean of observations at a given set of covariate values
- A new individual observation at a given set of covariate values.

The formulas for the standard errors for these kinds of predicted and fitted values are complicated and we omit them.

However, R can calculate these intervals easily using the command predict(...) function, so we can still talk about such predictions.

9 Testing for interaction

In multiple regression, two variables X_1 and X_2 are said to interact when the relationship between X_1 and the response variable Y depends on the values of the second variable X_2, and vice versa.

If X_1 and X_2 are the covariates between which an interaction is suspected, we incorporate the interaction between X_1 and X_2 by setting, for each $i \in 1, ..., n$,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$$

Then, the model can be written in two ways:

$$Y_i = \beta_0 + (\beta_1 + \beta_3 x_{i2}) x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

or

$$Y_i = \beta_0 + \beta_1 x_{i1} + (\beta_2 + \beta_3 x_{i1}) x_{i2} + \varepsilon_i$$

We refer to β_3 as the coefficient for the interaction between X_1 and X_2.

9.1 Procedure

1. Fit the model including the two covariates and the interaction.
2. Conduct a global F-test to assess whether any of the regression coefficients are different from zero, i.e., test the hypothesis
$$\mathcal{H}_0 \ : \ \beta_1 = \beta_2 = \beta_3 = 0$$
3. If this hypothesis is rejected, then test for an interaction by using a Student t-test to test the interaction null hypothesis $\mathcal{H}_0 \ : \ \beta_3 = 0$.
4. If the interaction null hypothesis is rejected, i.e. $\beta_3 \neq 0$, stop. Then, with this model, we can no longer interpret β_1 as the change in the mean of Y for a one-unit change in X_1, because that change now depends on β_3 and the value of X_2. We can only interpret the association of X_1 with the response Y depending on the value of X_2. Similarly, we can only interpret the association of X_2 with the response Y depending on the value of X_1.
5. Otherwise, re-fit the model without the interaction term to obtain estimates of β_1 and β_2 and interpret them in the usual way.

9.2 Interpretation

We can write the association of X_2 with the response variable, conditional on $X_1 = x$ as $\hat{\beta}_2 + \hat{\beta}_3 x$. So a one-unit increase in X_2 for a value of $X_1 = x$ will correspond to an estimated $\hat{\beta}_2 + \hat{\beta}_3 x$ increase in the mean price.

Multiple Linear Regression with Interactions and how to implement it in R

10 Qualitative variables in regression models

Qualitative variables like "Y/N" do not take numerical values. To use regression with qualitative variables, they must be assigned a numerical value. One way to do this is the following:

- set $Z = 1$ if severe covid
- set $Z = 0$ otherwise.

The variable Z can then be used as a covariate in the regression model. The group coded as $Z = 0$ is called the **reference group** for the analysis. An appropriate model would be

$$Y_i = \beta_0 + \beta_1 z_i + \varepsilon_i$$

where the qualitative variable with

- $Z = 1$ has a mean level of $\beta_0 + \beta_1$ and
- $Z = 0$ has a mean level of β_0.

In other words, β_1 measures the difference of means between the two.

$$\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0,$$

And also

$$\mathcal{H}_0 \ : \ \beta_1 = 0 \iff \mathcal{H}_0 \ : \ \mu_1 = \mu_0,$$

where μ_1 and μ_0 denote the true (population) means.

A Some Equations

Sum of squares

$$S_{XX} = \sum (x - \bar{x})^2$$
$$S_{YY} = \sum (y - \bar{y})^2$$
$$S_{XY} = \sum (x - \bar{x})(y - \bar{y})$$

Estimation

$$SSE = S_{YY} - \hat{\beta}_1 S_{XY} = S_{YY} - \frac{S_{XY}^2}{S_{XX}}$$

R

https://www.scribbr.com/statistics/anova-in-r/