LINEAR REGRESSION WITH A SINGLE-PREDICTOR

ETHAN LEVIEN

CONTENTS

The main subject of this course is linear regression models. In the simplest case where we have a single independent variable, x, a regression model for the relationship between Y and x is

$$y \sim \text{Normal}(a + bx, \sigma) \tag{1}$$

The variable x is called the **predictor**. Written another way, our model for Y is

$$y = a + bx + \xi, \quad \xi \sim \text{Normal}(0, \sigma). \tag{2}$$

Importantly, the standard linear regression model is a conditional model, in the sense that is tell us the distribution of our dependent variable, y, *conditioned* on x. This means that the distribution of x is not part of our model and in order to make predictions, the values of x need to be given to us. In order to emphasize this point, we sometimes write

$$y|x \sim \text{Normal}(a + bx, \sigma) \tag{3}$$

We are interested in the associated statistical inference problem, but first, let's assume a and b are known and think about some of the predictions we can make.

Example 1. *Suppose that we model the temperature y in degrees Fahrenheit on a given day as a function of the temperature on a previous day*

$$y \sim \text{Normal}(0.5x + 30, 3.) \tag{4}$$

If the temperature was 88 yesterday, what is the chance it is greater than 88 tomorrow? Use properties of the Normal distribution and confirm your result with simulations

Solution:
y is Normal with a mean and standard deviation of 70 and 3 respectively. From properties of the Normal distribution, the chance that $y >$ is 64.

```
> y = np.random.normal(70,5,1000)
> len(y[y>70])/len(y)
```

Example 2. *Suppose that we model the temperature y in degrees Fahrenheit on a given day as a function of the temperature on a previous day*

$$y \sim \text{Normal}(0.5x + 30, 1.) \tag{5}$$

If the temperature was 88 yesterday, what is the chance it is greater than 88 tomorrow?

Exercise 1. *Continuing with the temperature model from the previous question,*

`https://colab.research.google.com/drive/1H24olTdEAPp2xpxhMXhyvFaEzjjaiZnk#scrollTo=HdT4XpUZoVHk&line=1&uniqifier=1`

Date: April 2022.

2. Inference for linear regression models

Now suppose we have some data D consisting of x and y pairs:

$$D = \{(x_1, y_1), \ldots, (x_n, y_n)\} \tag{6}$$

One can show that the MLE estimates of β and α are

$$\hat{b}_{\text{MLE}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{7}$$

$$\hat{a}_{\text{MLE}} = \bar{y} - \hat{\beta}\bar{x} \tag{8}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of x and y respectively. This approach to estimating the regression slopes is also known as **least squares**, because it minimizes the squared error between the **residuals**

$$r_i = \hat{b}x_i + \hat{a} - y_i. \tag{9}$$

Exercise 2. *Discuss the relationship between the residuals and the noise term ξ_i, the difference between the data and the regression line. Are they equal? Why or why not?*

There is another parameter in our model, $\hat{\sigma}$. This also has an MLE. Recall that the estimator of a the standard deviation of a Normal distribution is

$$\sqrt{\frac{1}{n-1}\sum y_i^2} \tag{10}$$

If we want to estimate the standard error, we might expect to replace y_i with r_i (can you see why?); however, this does not account for the fact that we don't know $\hat{\alpha}_i$. This additional degree of freedom causes the Equation (**??**) to over estimate the variance. Correcting for this yields

$$\hat{\sigma} = \sqrt{\frac{1}{n-2}\sum y_i^2} \tag{11}$$

2.1. Linear regression with statsmodels.

We can perform a linear regression in stats models using the following code

```
> model= sm.OLS(y,X)
> results = model.fit()
```

This first command creates the "model" object, which is all the information about the data and the statistical model (linear regression). The second commend actually does the computations which give the results.[1] The results variable stores all the information about the fitted model, which we print with

```
>  print(results.summary())
```

This will output something like

```
>                             OLS Regression Results
> ==============================================================================
> Dep. Variable:                      y   R-squared:                       0.132
> Model:                            OLS   Adj. R-squared:                  0.123
> Method:                 Least Squares   F-statistic:                     14.86
> Date:                Mon, 13 Sep 2021   Prob (F-statistic):           0.000207
> Time:                        21:28:39   Log-Likelihood:                 89.756
> No. Observations:                 100   AIC:                            -175.5
> Df Residuals:                      98   BIC:                            -170.3
> Df Model:                           1
> Covariance Type:            nonrobust
> ==============================================================================
>                  coef    std err          t      P>|t|      [0.025      0.975]
> ------------------------------------------------------------------------------
> const          1.9925      0.010    199.914      0.000       1.973       2.012
> x1             0.1712      0.044      3.855      0.000       0.083       0.259
> ==============================================================================
> Omnibus:                        2.592   Durbin-Watson:                   1.959
> Prob(Omnibus):                  0.274   Jarque-Bera (JB):                2.382
> Skew:                          -0.040   Prob(JB):                        0.304
> Kurtosis:                       3.752   Cond. No.                         4.46
> ==============================================================================
```

[1] The function OLS uses maximum likelihood as described in Chapter 8 of the textbook. This is constrast to the Bayesian methods used in Chapter's 6 and 8. We will learn how to approach this problem from a Bayesian perspective later on, but for now, note that the results are slightly different.

Example 3. *Generate simulated data from the model*

$$y \sim \text{Normal}(4x + 5, 5) \tag{12}$$

and perform a linear regression on the fitted model.

3. Correlation and standardization

We start with an exercise in rewriting the estimator of β: Let $\hat{\sigma}_x^2$ and $\hat{\sigma}_y^2$ be the estimators of the variance in x and y. Remember that even though we don't have a model for x, we can still compute the standard deviation. The sample correlation of two samples x and y is given by

$$r_{x,y}^2 = \frac{1}{n-2}\sum(x_i - \bar{x})(y_i - \bar{y}) \tag{13}$$

Note that $\hat{\beta} = r_{x,y}/\hat{\sigma}_x^2$. Sometimes it is useful to *standardize* the variables before performing a regression, meaning we transform them into standard Normal random variables.

Exercise 3. *Show that for a sample of a random variable $x_1, x_2, \ldots, x_n$*

$$z_i = \frac{x_i - \bar{x}}{\hat{\sigma}_x^2} \tag{14}$$

If we perform a regression on the standardized variables, then

$$\hat{\beta} \tag{15}$$

The quantity ρ is called the *correlation coefficient*. We will take

Exercise 4. *Is the correlation coefficient symmetric; that is, i*

4. Hypothesis testing for regression models

We've already discussed p-values and mentioned some potential reasons to avoid using them. However, they play a central role and statistics and we must therefore understand them in the context of regression.

5. Additional exercises

Exercise 5. *Using the code above, write a function which does the following: First, it generates data from a simulated linear regression with slope a, intercept b, measurment noise sigma and n data points. It then performs a linear regression on the simulated data and outputs the estimated slope, intercept, the R^2 value and the p-value for the slope. You can generate the x values by copying the code at the beginning of this section and changing n.*

Exercise 6. *Using your function, make a plot of the p-value vs. n using the values of a, b and sigma from above. Make the same plot with $a = 1.0$, $b = 2$ and $\sigma = 0.1$. Now repeat this but plotting R^2 instead of the p-values. Can you explain why the behavoir of these plots makes sense? Hint: It may be helpful to plot the results on a log scale using 'ax.semilogy'.*

Exercise 7. *Generate data such that performing a linear regression results in a (a) A very small p-value and an R^2 value close to 1. (b) A large p value and a very small R^2 value. (c) A small p-value and a small R^2 value.*