Lab 11: linear regression and Inference

Your Name and Student ID

Today's Date

Review

Let's assume that we take a random sample of size n from some population where each observation consists of a response variable Y and an explanatory variable X. It may be reasonable to assume that each observation's response value is generated in a linear fashion based on the observation's explanatory variable's value. Although there are many ways to formulate a such a linear model, we're familiar with the line of best fit for the estimated response:

$$\hat{Y} = a + bX_i$$

To write an equation for the observed value for y, we simply add the residual which denotes the deviation between the line of best fit and the actual value.

$$Y = a + bX_i + e$$

In these equations, a is the population intercept, b is the population slope and e is the residual. In other words, when we assume this model, we assume that each observation response variable is generated based on the value of its explanatory variable, some error term, a common intercept parameter and a common slope parameter. However, we do not know the values of a or b; we must estimate them using our sample. The random error, e, is also unknown. We can approximate it by $\hat{e}$. We compute the estimated residuals as follows: $\hat{e} = Y - \hat{Y}$. Note that in the most recent lecture, we denoted the residuals by r and their estimate by $\hat{r}$.

Once we have estimated the linear model parameters, denoted $\hat{a}$ and $\hat{b}$ we can calculate an estimated response value for each observation in our sample using the following equation:

$$\hat{Y} = \hat{a} + \hat{b}X$$

We can also write the equation above as:

$$Y = \hat{a} + \hat{b}X + \hat{e}$$

Recall that $\hat{a}$, $\hat{b}$ and $\hat{e}$ are statistics based on our random sample; they will vary from sample to sample. If we are interested in measuring the effect that the explanatory variable, X, has on the response variable, Y, then we are likely interested in the the parameter b. Unfortunately, b is hidden from us, all that we have its estimate $\hat{b}$. Luckily, if our data satisfies certain assumptions, we can use the $\hat{b}$ to make inferences on b. These assumptions are listed below:

1. The relationship between X and Y is linear in the population.
2. The response variable varies Normally (i.e. following a Normal distribution) around the line of best fit. Equivalently, the residuals are identically Normally distributed for each observation. Note that this does not mean that the response variable must be Normally distributed.
3. The standard deviations of the response variable are identical for all values of the explanatory variable.
4. Each observation is independent.

In this discussion, we will explore the methods for validating these assumptions through an example.

Part I: Checking assumptions for linear regression

Boston data on median household value and air pollution

We are examing a data set used to predict housing prices in the area around Boston (Harrison, D. and Rubinfeld, 1978). We wish to examine specifically the association of the measure of housing price (`medv`, median value of owner-occupied homes in the $1000s) and a measure of air pollution (`nox`, nitrogen oxides concentration, parts per 10 miillion). The data frame (Boston) is contained in another package (MASS), which we load below.

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
##   lstat medv
## 1  4.98 24.0
## 2  9.14 21.6
## 3  4.03 34.7
## 4  2.94 33.4
## 5  5.33 36.2
## 6  5.21 28.7
```

1. Perform and summarize the results of a linear regression of `medv` versus `nox` using Boston data. Be careful about which variable is explanatory and which is response!

2. Recall that before we only interpreted the estimate for the intercept and slope parameter. Interpret the slope parameter for `nox`. Notice also that the other columns are `std.error`, `statistic`, and `p-value` – these should remind you things we've learned during Part III of the course on inference. They correspond to the hypothesis test with the null hypothesis that the parameter is equal to 0. Thus, how would you interpret the p-value for `nox`? (We will cover this in detail in class on Friday.)

Write your answer here.

3. Use `glance()` to look at the r^2 value for this model. Does `nox` explain alot of the variance in median household value? Would you expect it to?

Write your answer here.

4. Check the assumptions required for the simple linear model using the plots shown during Wednesday's lecture. Note that to make these plots you first need to fit the linear model and then use the `augment()` function from the broom package to store the residuals and fitted values into a new data frame. Try to update the plotting code from class for your data set. The GSIs can help you if you get stuck!

The hardest one to make is likely the boxplots because the data first needs to be reshaped to make the plots. The reshaping code is the `gather()` function that you see in the slides/Rmd file for making these plots. Here is a helpful R explainer for how gather() works: https://twitter.com/WeAreRLadies/status/1059520693857996800. Basically, we need to gather the observed y values and the residuals by stacking them into one variable so that we can make two box plots side-by-side. Below, we include the `gather()` code for you since it is a bit tricky. You need to use the resulting data frame to make your box plots.

5. What do you think about the assumption plots? They are a bit messier than those shown in class, and reflect what we often see "in the real world".

Write your answer here.

Lab Conclusion (make sure to read this and understand it)**

From this exercise, we can conclude that there is a negative association between air pollution and median household value. An increase in Nitrogen Oxide of 1 parts per million (PPM) is associated with a decrease in median household value of $33,900 (see `help(Boston)` to remind yourself of the units for `nox` and `medv`). Note that this "increase of 1 unit" is wider the range of the scatter plot, so we should modify it to talk about an increase of 0.1 unit in `nox`. That is, an increase in Nitrogen Oxide of 0.1 PPM is associated with a decrease in median household value of $3,390. This you can see when you look at the scatter plot of the data and the line of best fit. – Look at going from 0.5 to 0.6 on the x axis and see how the model predicted y going from ~$25k to ~$22k.

Part II (Inference in Regression)

Boston data on median household value and distance to employment centers

We are examing a data set used to predict housing prices in the area around Boston (Harrison, D. and Rubinfeld, 1978). We wish to examine specifically the association of the measure of housing price (`medv`, median value of owner-occupied homes in the $1000s) and a measure of adjacency to employment (a weighted distance, roughly in miles). The data frame (Boston) is contained in another package (MASS), which we load below.

1. Perform and summarize the results of a linear regression of `medv` (median value of owner-occupied homes in $1000s) and `dis` (weighted mean of distances to five Boston employment centres) using Boston data. Be careful about which variable is explanatory and which is response!
2. Interpret the slope parameter and p-value from the table. What null and alternative hypotheses does this p-value refer to?
3. Derive a 95% CI for this slope parameter. In your opinion, would you expect the direction of this relationship to hold if the data were collected today?

Write your answer here.

4. Use a function to look at the r-squared value for this model. Does `dis` explain alot of the variance in median household value? Would you expect it to?

Write your answer here.

5. Back to the fit of the model of `medv` vs. `dis`. Make a plot with the raw data points, the fitted line from the simple linear regression model (only containing `medv` and `dis`) and also add a line with a slope of 0. You can have that line cross the y axis at the average value of `medv` to vertically bisect the data points.
6. For you, does the plot raise any concerns about the assumptions of the linear regression you just performed? What other plots might you do to explore the fit?

Write your answer here.

(Optional) Pointwise Confidence Intervals and Multiple Testing

As you learned in lecture, there are two types of confidence intervals applicable to estimating a point on the plot which are related to whether one is predicting the population average among individuals with $X = x$ (**mean response**) or whether one is predicting the actual Y for a particular individual (**single observation**). For this assignment, we will concentrate on the confidence interval for the mean response. We do so because it is rare to use statistical models in public health as forecasting models (predicting an individual's health in the future) and more common to use them to estimate population-level changes (how does the mean health change in a population as we change exposure). However, as precision medicine becomes more of a reality and the models accurately predict health (i.e., have high R^2's), then statistical forecasting may become more common in our field.

7. Calculate four 95% confidence intervals, one at each `dis` value: 2.5, 5.0, 7.5, and 10.0 miles. Add the four CIs to a scatter plot of the data (along with the line of best fit):
8. Interpret the pointwise 95% of the median house price when distance = 10.

Write your answer here.

9. Why do you think the CI's get wider as the `dis` gets larger?

Write your answer here.

10. Add a prediction interval at x = 2.5. Why is is wider or more narrow?

Write your answer here.

Submission

Please submit your lab *directly* to Gradescope. You can do this by knitting your file and downloading the PDF to your computer. Then navigate to Gradescope.com to submit your assignment. Here is a tutorial if you need help: https://www.gradescope.com/get_started. Scroll down on that page to “For students:submitting homework”.