DS 100/200: Principles and Techniques of Data Science **Date: April 17, 2020**

Discussion #11 Solutions

Name:

Residuals

1. (a) We fit a simple linear regression to our data $(x_i, y_i), i = 1, 2, 3$, where x_i is the independent variable and y_i is the dependent variable. Our regression line is of the form $\hat{y} = \hat{a} + \hat{b}x$. Suppose we plot the relationship between the residuals of the model and the $\hat{y}s$, and find that there is a curve. What does this tell us about our model?

   ☐ A. The relationship between our dependent and independent variables is well represented by a line.

   ☐ B. The accuracy of the regression line varies with the size of the dependent variable.

   ☐ C. The variables need to be transformed, or additional independent variables are needed.

   Solution:
   If we see a curve in our residual plot, then the relationship is not well represented by a line. Either more independent variables are needed, or transformations of the current variables are necessary.

   (b) Which of the following are useful properties of residuals in an ordinary least squares regression (with an intercept) where the dependent and independent variables are vectors?

   ☐ A. The average of the residuals is 0.

   ☐ B. The inner product of the fitted values and the residuals is always positive.

   ☐ C. The residuals and the independent variable are orthogonal.

   Solution:
   Properties of residuals.

Inference

2. We can use the bootstrap to carry out inference on the slope of a simple linear regression. Recall that a simple linear regression model is defined as follows:

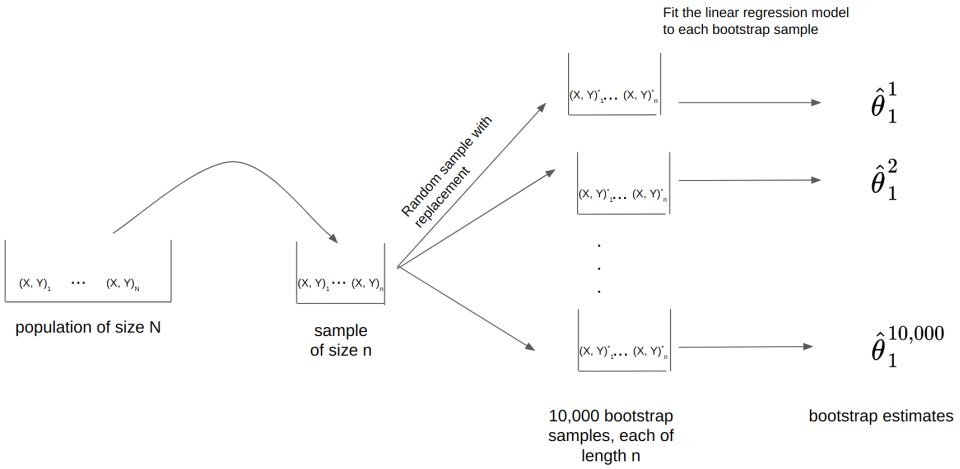
$$Y = \mathbf{X}\theta + \epsilon = \theta_0 + \theta_1 x + \epsilon$$

where Y is the response vector, X is the design matrix with an added intercept column, and ϵ is the noise vector. Using the data to estimate the intercept and the slope, we arrive at the following equation:

$$f_{\hat{\theta}}(x) = \hat{\theta}_0 + \hat{\theta}_1 x$$

(a) Using a box model, describe the process of computing the 95% confidence interval of $\hat{\theta}$.

Solution:



Given the bootstrap estimates, we can estimate the sampling distribution of θ_1. The 95% confidence interval is approximated by the 2.5^{th} percentile of the bootstrap estimates distribution, and the upper-endpoint is approximated by the 97.5^{th} percentile.

Depending on the size of N and n, the original sample can be a simple random sample with or without replacement. If the sample is much smaller than the size of the population, then sampling without replacement is acceptable since observations are approximately independent. However, if the size of the sample is large, sampling should be performed with replacement to ensure that each observation in the sample is drawn independently.

(b) Now that we have some intuition for how the bootstrap works in a simple linear regression, let's think about how we might implement this for a multivariate linear regression. Suppose we wish to fit a model of the following form:

$$f_{\hat{\theta}}(x) = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \cdots + \hat{\theta}_p x_p$$

and we would like to generate confidence intervals around our estimates of θ. Outline in pseudo-code a non-parametric bootstrap based approach to estimate the 95% confidence interval around each θ_i. Assume there are n data points.

Solution:
$theta\_hats$ = []

For $i = 1$ to $num\_replicates$:

$bootstrap\_sample = SampleWithReplacement(data, n)$

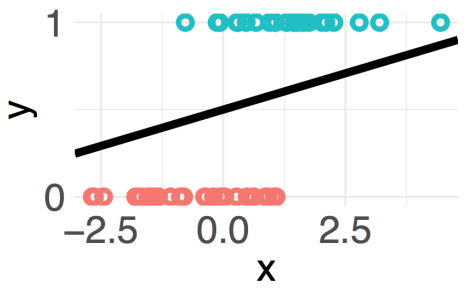
$theta\_hat = LinearModel.fit(bootstrap\_sample).coefficients$

$theta\_hats.append(theta\_hat)$

$.025\_CI = percentile(theta\_hats, .025)$
$.975\_CI = percentile(theta\_hats, .975)$

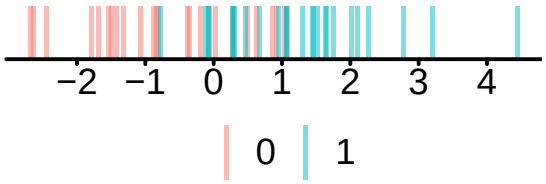
Logistic Regression

3. Your friend argues that the data are linearly separable by drawing the line on the following plot of the data.



(a) Argue whether or not your friend is correct. Note: this question refers to a binary classification problem with a single feature.

Solution: The scatter plot of x against y isn't the graph you should be looking at. The more salient plot would be the 1D representation of the features colored by class labels.



From this plot, it's clear that we can't draw a point on the axis that separates the data.

(b) Suppose you use gradient descent to train a logistic regression model on two design matrices $\mathbb{X}_a$ and $\mathbb{X}_b$ and use some stopping criterion (The maximum of the absolute values of the components in the calculated gradient is smaller than some threshold T). After training, you find that the training accuracy for $\mathbb{X}_a$ is 100% and the training accuracy for $\mathbb{X}_b$

is 98%. What can you say about whether the data is linearly separable for the two design matrices?

Solution: We can say that the data is linearly separable using $\mathbb{X}_a$ because we have found a line that separates the two classes. We cannot say anything about $\mathbb{X}_b$ because we may have chosen the wrong threshold, or alternatively because gradient descent may not have found the global minimum.

4. Suppose we are given the following dataset, with two features ($\mathbb{X}_1$ and $\mathbb{X}_2$) and one response variable (y).

$\mathbb{X}_1$	$\mathbb{X}_2$	y
1	1	0
1	-1	1

Here, **x** corresponds to a single row of our data matrix, not including the y column. For instance, $\mathbf{x}_1 = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$

You run an algorithm to fit a model for the probability of $Y = 1$ given **x**:

$$\mathbb{P}\left(Y = 1 \mid \mathbf{x}\right) = \sigma(\mathbf{x}^T \beta)$$

where

$$\sigma(t) = \frac{1}{1 + \exp(-t)}$$

Your algorithm returns $\hat{\beta} = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}^T$

(a) Are the data linearly separable? If so, write the equation of a hyperplane that separates the two classes.

Solution: Yes, the line $\mathbb{X}_2 = 0$ separates the data in feature space.

(b) Recall that the empirical risk for $\hat{\beta} = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}^T$ and the two observations above is $\frac{1}{2}\log(2 + 2e^{-1})$. Does this fitted model minimize cross-entropy loss?

Solution: No.

Our dataset is linearly separable, which means that the absolute values of components in $\vec{\beta}$ diverge to ∞, and the optimal cross entropy loss is 0. A cross entropy loss of 0 can never actually be achieved (remember, $\sigma(t)$ never outputs exactly 0 or 1), but gradient descent will bring this value arbitrarily close to 0.

No single value of $\vec{\beta}$ will "minimize" cross entropy loss, because we can always pick a value of $\vec{\beta}$ that has an even smaller loss.

In order to avoid our absolute values of the weights diverging to ∞, we can regularize our cross entropy loss.

Multicollinearity

5. In a large class, the instructor decides to build a linear model to predict final exam scores based on the scores of the past two midterms. Using scores from the previous year, the instructor derives the following model from the least-squares regression equation:

$$\hat{Y}(x) = 1.5x_1 - 0.5x_2$$

where $\hat{Y}(x)$ are the predicted final scores, x_1 are the midterm 1 scores, and x_2 are the midterm 2 scores.

(a) Does the negative coefficient of x_2 imply that Midterm 2 scores and final exam scores are negatively correlated? Why or why not?

> **Solution:** It's possible that the two are negatively correlated, but our intuition tells us that there's likely a strong correlation between midterm 1 scores and midterm 2 scores and that our model is in danger of collinearity.

(b) The instructor, confused by the weights of his linear regression model, computes the 95% confidence intervals for the two coefficients.

Confidence interval for θ_1 : $[-0.32, 1.88]$
Confidence interval for θ_2 : $[-0.5, 1.16]$

Are θ_1 and θ_2 significantly different from 0? If both coefficients are not significantly different from 0, does that suggest that both midterm 1 and midterm 2 are uncorrelated with the final?

> **Solution:** Because 0 is in both of the 95% confidence intervals, θ_1 and θ_2 are not significantly different from 0.
>
> We cannot make any conclusions about the correlation between the midterms and the final because it is possible that the midterms are collinear.
>
> Note: In the Lecture 22 Snowy Plover example, we found that dropping collinear features can change the significance of our coefficients.