Linear Regression Models
Final Exam

December 21 2020

If you have any questions, please email to arnab.auddy@columbia.edu and/or xs2427@columbia.edu

1. (20 points) The following regression output was obtained using the city-economy data set. Recall that for each of 366 cities in the US, this records the city's per-capita gross metropolitan product, in dollars per person per year, and its population.

```
x = log10(pop) y = pcgmp

out = lm(y ~ x)
summary(out)

### log10 computes log to the base 10. For example, log10(100) = 2.
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##   Min    1Q  Median 3Q   Max
## -21572 -4765 -1016 3686 40207
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -23306       4957    -4.7  3.7e-06
## x              10246        900    11.4  < 2e-16
##
## Residual standard error: 7930 on 364 degrees of freedom
## Multiple R-squared: 0.263,Adjusted R-squared: 0.26
## F-statistic: 130 on 1 and 364 DF, p-value: <2e-16
```

For the following questions, explain clearly which parts of the output are the basis for your answers.

1.1. (4 points) What is the predictor variable? What is the response variable? Which variables were transformed, and how?

1.2. (3 points) Write the equation for the estimated conditional mean function; use numerical values rather than symbols like $\widehat{\beta}_0$.

1.3. (3 points) According to the estimated model, what is the average per-capita gross metropolitan product of cities with a population of one million people? Of cities with a population of two hundred thousand people? Do these numbers seem reasonable?

1.4. (2 points) Based on the estimated coefficients, can you give an estimate of $\mathbb{E}[Y|X=0]$? If yes, what is it (and show your work); if not, explain why not.

1.5. (2 points) Give a 95% confidence interval for β_1, assuming all the model assumptions hold.

1.6. (2 points) Give an estimation of σ^2.

1.7. (2 points) Can you find the sample variance of the variable `pop` from the information in the output? If so, what is it? If not, explain.

1.8. (2 points) Which part (or parts) of the output (if any) tests the assumption that the relationship between the predictor variable and the response variable is linear?

2. (15 points) Suppose we have a regression fit with p predictors as

$$y_i = \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i \qquad \text{for } i = 1, \ldots, n.$$

Assume that the sum of all Y and x_j values satisfy $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_{ji} = 0$.

We now have a new predictor x_{p+1} which is orthogonal to all the columns of X. That is, $x_j^T x_{p+1} = 0$ for $j = 1, \ldots, p$. Moreover, $\sum_{i=1}^{n} x_{(p+1),i} = 0$.

Prove that the least squares estimate $\hat{\beta}_{p+1}$ in the new regression (involving all $p+1$ predictors) is same as the univariate regression coefficient of regressing Y on x_{p+1}.

Hint: Consider any matrix X and a new one $X' = [X\, x_{p+1}]$. Then the orthogonal projection matrices onto the column spaces of X and X' satisfy

$$P_{X'} = P_X + \frac{1}{x_{p+1}^T (I - P_X) x_{p+1}} \cdot (I - P_X) x_{p+1} x_{p+1}^T (I - P_X).$$

3. (20 points) Consider the following algorithm for leave one out cross validation (LOOCV).

3.1. (5 points) In class, we have mentioned LOOCV several times. What is the main use of LOOCV and why do we need it? Compare to K-fold cross validation, why do people prefer LOOCV when the number of samples is large?

```
for $i = 1$ to n do
    Compute the datasets $(X_{-i}, Y_{-i})$ by removing the $i^{\text{th}}$ observations.
    Regress $Y_{-i}$ on $X_{-i}$. Let the coefficients be $\beta_{-i}$.
    Find the best prediction for $Y_i$ as $\hat{Y}_{i,(-i)} := x_i^T \beta_{-i}$.
end for
```

3.2. (3 points) Count how many matrix inversions you have to compute in the above algorithm.

3.3. (5 points) Using results stated in class/textbook (mention which ones you use, without proof) show that we can write

$$\hat{Y}_i = \gamma_i Y_i + (1 - \gamma_i)\hat{Y}_{i,(-i)}.$$

Obtain an analytical expression for γ_i in terms of the design matrix $\mathbf{X}$.

3.4. (7 points) Based on the above, suggest a faster algorithm for performing LOOCV. How much have you improved from part 3.1?

4. (15 points) It has been claimed that gross metropolitan products show a simple quantitative regularity, called "supra-linear power-law scaling". If Y is the gross metropolitan product in dollars, and N is the population of the city, then, the claim goes,

$$Y \approx cN^{r-1}$$

where the exponent $r > 1$ and the scale factor $c > 0$. (If this model holds with an exponent $r < 1$, there is said to be "sub-linear scaling".) Assume that we observe $Y_i = cN^{r-1}\exp(Z_i)$, where $Z_i \overset{iid}{\sim} N(0, \sigma^2)$.

4.1. (2 points) Does it make sense to run a linear regression of Y on N? What should we do instead?

4.2. (3 points) A possible idea is to regress $\log Y$ on $\log N$. Suppose the intercept and slope estimates are b_0 and b_1 respectively. What will your estimates be for c and r?

4.3. (4 points) We have a 95% confidence interval of $[-2.31,\ 4.16]$ for the intercept in the regression in part 4.2. What should the corresponding interval be for c?

4.4. (2 points) We want to test whether a city shows supra-linear or sub-linear scaling. Construct a suitable hypothesis test for this, mentioning the null and alternative.

4.5. (4 points) Upon running the regression in 4.2 using data from 12 cities, we find $b_1 = 0.36$ and $se(\hat{\beta}_1) = 0.2$. You are given the following table of critical values from the t distribution:

df	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.025$
10	1.372	1.812	2.228
11	1.363	1.795	2.201
12	1.356	1.782	2.179

Table 1: Critical values of the t distribution

What is your conclusion on the hypothesis in 4.4?

5. (15 points) Consider a logistic regression model, that is

$$\mathbb{P}(Y=1|X=x)=\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}=1-\mathbb{P}(Y=0|X=x).$$

After fitting the regression, we predict the class as $\hat{\mu}(x)$, which can be either 0 or 1.

5.1. (5 points) Justify that $\mathbb{P}(Y\neq\hat{\mu}(x))=\mathbb{E}(Y-\hat{\mu}(x))^2$.

5.2. (5 points) Next, show that $\mathbb{E}[(Y-\hat{\mu})^2|X=x]=\mathbb{P}(Y=1|X=x)(1-2\hat{\mu}(x))+\hat{\mu}^2(x)$.

5.3. (5 points) Calculate the misclassification errors in the different cases (Hint: make a contingency table) and justify the following choices:

- If $\mathbb{P}(Y=1|X=x)>0.5$, we should have $\hat{\mu}(x)=1$.
- If $\mathbb{P}(Y=1|X=x)<0.5$, we should have $\hat{\mu}(x)=0$.
- If $\mathbb{P}(Y=1|X=x)=0.5$, both predictions are equally risky.

6. (15 points) Consider the linear relationship $Y=x\beta+\epsilon$, where $\epsilon\sim\mathcal{N}\left(0,\sigma^2 I_n\right)$, $\beta\in\mathbb{R}^p$.

6.1. (2 points) Give an intuitive justification of the Mallows C_p criterion.

6.2. (3 points) Describe the AIC and BIC methods of model selection.

6.3. (2 points) How are AIC and BIC different? Mention which one you should use depending on what the true model is.

6.4. (3 points) Now we have a new criterion $D=2\sum_{i=1}^{p}|\beta_i|-2\log L$, where L is the likelihood of the model. Write down the expression for D and AIC in the above mentioned specific model and compare the difference of them.

6.5. (5 points) Compare D with the loss function given by Lasso, and conclude the difference between the model selected by the AIC and with the D criterion.