Final Exam

Aiden Kenny
STAT GR5205: Linear Regression Models
Columbia University

December 21, 2020

Question 1

Let Y denote per-capita gross metropolitan product (GMP), in dollars per person per year, and X denote population, in people. The realized values of these random variables are respectively given by the n-vectors $\mathbf{y}$ and $\mathbf{x}$, where $n = 366$.

1. The predictor variable is given by $Z \coloneqq \log_{10} X$, and the response is Y. We can see that the population is being transformed by taking the logarithm (with base 10).

2. Our estimated model is given by

$$\mathbb{E}(Y \mid x) = -23306 + 10246 \log_{10} x. \tag{1}$$

3. We have $\mathbb{E}(Y \mid 1,000,000) = 38170$ and $\mathbb{E}(Y \mid 200,000) = 31008.35$. These answers make sense, a city with a larger population will have a higher GMP per-capita.

```
-23306 + 10246 * log(c(1000000, 200000), 10)
```

4. We cannot give an estimate of $\mathbb{E}(Y \mid 0)$ because $\log_{10} 0$ is undefined.

5. A 95% confidence interval for β_1, denoted as $\mathcal{I}_{\beta_1}$, is

$$\mathcal{I}_{\beta_1} = \Big(\hat{\beta}_1 - t \cdot \operatorname{se}(\hat{\beta}_1), \hat{\beta}_1 + t \cdot \operatorname{se}(\hat{\beta}_1)\Big) = \big(10246 - 1.967 \cdot 900, 10246 + 1.967 \cdot 900\big) = \big(8475.7, 12016.3\big). \tag{2}$$

   The values $\hat{\beta}_1 = 10246$ and $\operatorname{se}(\hat{\beta}_1) = 900$ can be found in the `R` output, and the value $t = T_{364}^{-1}(0.975) = 1.967$ can be found using the `qt()` function in `R`.

```
qt(0.975, 364)
```

```
10246 + 1.967 * 900 * c(-1, 1)
```

6. From the `##Residual standard error` section, we have $\hat{\sigma}^2 = 7930^2/364 = 172760.7$.

```
7930^2 / 364
```

7. You cannot find the sample variance of X from the `R` output. We are never considering the value of $\operatorname{Var}(Z)$ when constructing the model because we are never treating Z as a random variable. We instead are treating it as a set of fixed values $\mathbf{z}$, either observed before or after the model's design is chosen. When we estimate σ^2 in the linear model, we are estimating $\operatorname{Var}(\epsilon)$, the residuals of the model. And since we cannot make any inferences about $\operatorname{Var}(Z)$, we cannot make any inferences about $\operatorname{Var}(X)$ either.

8. There are multiple components of the `R` output that test the hypothesis $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$. Remember, the output is testing the hypothesis that Y and Z have a linear relationship, *not* Y and X. There are two tests that `R` runs when using the `lm()` function: the t test and the ANOVA test. The p-palue for the t test is found in the right-most column, `Pr(>|t|)`, of the `##Coefficients` section, and is given by `<2e-16`. The p-value for the ANVOA test is found in the last entry in the output, in the `##F-statistic` section, and is also given by `<2e-16` (`R` will estimate the value if it is too small). In both cases, we reject H_0, and it seems that there is indeed a linear relationship between Y and Z $(= \log_{10} X)$.

Question 2

Suppose we have observed n observations of p predictors, given by $\mathbf{x}_1, \ldots, \mathbf{x}_p$, and an observed response $\mathbf{y}$. Let $\mathbf{X} \coloneqq \begin{bmatrix}\mathbf{x}_1 \cdots \mathbf{x}_p\end{bmatrix}$ be a matrix where the jth column is $\mathbf{x}_j$. Here we also assume that the data has been centered, so each predictor and the response has a mean of zero (this is common to do before fitting a model). We fit a regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, and our estimated coefficients are given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

Now, suppose we have n observations of a new predictor $\mathbf{z}$, which was not used at all when determining $\hat{\boldsymbol{\beta}}$. It turns out that this new predictor is orthogonal to each of the previous p predictors, i.e. $\mathbf{x}_j^T\mathbf{z} = 0$ for all j, and so $\mathbf{X}^T\mathbf{z} = \mathbf{0}$. If we want to fit a new linear model that includes $\mathbf{z}$, we can use an alternate matrix $\widetilde{\mathbf{X}} \coloneqq \begin{bmatrix}\mathbf{X} & \mathbf{z}\end{bmatrix}$ and fit the model $\mathbf{y} = \widetilde{\mathbf{X}}\boldsymbol{\beta}$; our estimated coefficient will be given by $\hat{\boldsymbol{\beta}} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}\mathbf{y}$. As we will soon see, the estimated coefficients for the original p predictors is exactly the same as they were in the origianl model (denoted as $\hat{\boldsymbol{\beta}}_0$), and the estimated coefficient for $\mathbf{z}$ would be the same if a model was fit using *only* this new predictor! The key reason for both of these results is that the new predictor is orthogonal to each of the previous ones. To see this, we first observe that

$$\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} = \begin{bmatrix}\mathbf{X}^T \\ \mathbf{z}^T\end{bmatrix}\begin{bmatrix}\mathbf{X} & \mathbf{x}\end{bmatrix} = \begin{bmatrix}\mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{z} \\ \mathbf{z}^T\mathbf{X} & \mathbf{z}^T\mathbf{z}\end{bmatrix} = \begin{bmatrix}\mathbf{X}^T\mathbf{X} & \mathbf{0} \\ \mathbf{0}^T & \mathbf{z}^T\mathbf{z}\end{bmatrix}.$$

Using the formula for matrix inversion for block matrices (see `here`), we have

$$(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1} = \begin{bmatrix}\mathbf{X}^T\mathbf{X} & \mathbf{0} \\ \mathbf{0}^T & 1/\mathbf{z}^T\mathbf{z}\end{bmatrix},$$

and so the estimated coefficients are given by

$$\hat{\boldsymbol{\beta}} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}\mathbf{y} = \begin{bmatrix}\mathbf{X}^T\mathbf{X} & \mathbf{0} \\ \mathbf{0}^T & 1/\mathbf{z}^T\mathbf{z}\end{bmatrix}\begin{bmatrix}\mathbf{X}^T \\ \mathbf{z}^T\end{bmatrix}\mathbf{y} = \begin{bmatrix}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ \mathbf{z}^T\mathbf{y}/\mathbf{z}^T\mathbf{z}\end{bmatrix} = \left(\hat{\boldsymbol{\beta}}_0, \frac{\mathbf{z}^T\mathbf{y}}{\mathbf{z}^T\mathbf{z}}\right)^T. \tag{3}$$

This result says that the first p estimated coefficients are given by $\hat{\boldsymbol{\beta}}_0$, while the new predictor's estimated coefficient is given by the value $\mathbf{z}^T\mathbf{y}/\mathbf{z}^T\mathbf{z}$. If we fit a simple linear model using only the new predictor, $\mathbf{y} = \beta\mathbf{z}$, the estimated coefficient is given by $\hat{\beta} = (\mathbf{z}^T\mathbf{z})^{-1}\mathbf{z}^T\mathbf{y} = \mathbf{z}^T\mathbf{y}/\mathbf{z}^T\mathbf{z}$.

The two main points of this question are that adding an orthogonal variable to a linear model does not change the value of the estimated coefficients of the previous variables, and obtaining the estimated coefficient for the new variable is as easy as computing two inner products. In fact, these two ideas are a possible method for estimating $\boldsymbol{\beta}$ for any linear model. One would first have to orthogonalize each of the variables to be used, and the estimated coefficient for the jth variable is then given by $\mathbf{x}_j^T\mathbf{y}/\mathbf{x}_j^T\mathbf{x}_j$ (see chapter 3 of `this book` for more info).

Question 3

1. Leave-one-out cross-validation (LOOCV) is a resampling method used to better estimate the effectiveness of a statistical model. We want to see how effective a model is at making predictions on data points not present when the model is being fit. Instead of having to collect a separate data set, we remove a single observation from the data set, fit the model with the remaining data, and then determine the error for that single point. Doing this for each point and taking the average of each error gives us our estiamted test error, given by

   $$\text{CVE} = \frac{1}{n}\sum_{i=1}^{n} d_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i,(-i)})^2,$$

   where d_i is the ith deleted residual. The idea of k-fold cross-validation (KCV) is similar, instead of leaving out one data point, we leave out n/k data points, and repeat the procedure k times to estimate the test error. In fact, LOOCV is a special case of KCV when $k = n$. When the number of samples is large, it is generally more feasible to use KCV (usually with $k = 5$ or $k = 10$) instead of LOOCV, since it is computationally faster and has a smaller variance. However, in the special case of fitting a linear model (the scenario in this problem), we can actually use a more efficient algorithm that drastically reduces the computational burden, so in this case it would be favorable to use LOOCV.

2. Since we are fitting a linear model n times, there will be n matrix inversions to compute.

3. Kutner etal, we have two equal formulations for the ith deleted residual: $d_i = y_i - \hat{y}_{i,(-i)}$, and $d_i = e_i/(1-h_{ii})$, where $e_i = y_i - \hat{y}_i$ is the ith residual from the model fitted with all observations and h_{ii} is the ith diagonal element of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Setting these equal to each other and solving for $\hat{y}_i$ gives us

$$\hat{y}_i = h_{ii}y_i + (1-h_{ii})\hat{y}_{i,(-i)}, \tag{4}$$

and so $\gamma_i = h_{ii} = [\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]_{ii}$.

4. Solving for $\hat{y}_{i,(-i)}$ gives us $\hat{y}_{i,(-i)} = (\hat{y}_i - h_{ii}y_i)/(1-h_{ii})$. This means that we can calculate every $\hat{y}_{i,(-i)}$ using information from the model fitted using every observation. The CVE is then given by

$$\text{CVE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{\hat{y}_i - h_{ii}y_i}{1-h_{ii}}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_{ii}}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2.$$

In order to obtain $\mathbf{e}$ and $\mathbf{H}$, *we only need to fit a single model using all of the predictors*! This is a drastic improvement from the original algorithm where we had to fit n different models.

Question 4

Let Y and N denote a city's GMP in dollars and population, respectively (we are not treating N as a random variable), and suppose we observe $Y = cN^{r-1}\exp(\epsilon)$, where $\epsilon \sim \text{N}(0, \sigma^2)$. Here c is a scaling factor and r is the scaling exponent. If $r > 1$ then the city has supra-linear scaling, if $r = 1$ then the city has linear scaling, and if $r < 1$ then the city has sub-linear scaling.

1. We cannot fit a linear model of Y with respect to N, since Y is not linear with respect to the model parameters. We can, however, take the logarithm of both sides to get

$$\log Y = \log c + (r-1)\log N + \epsilon, \tag{5}$$

and can now regress Y on $\log N$, since it is linear with respect to the new parameters $\beta_0 = \log c$ and $\beta_1 = r - 1$.

2. If our estimated coefficients are given by $\hat{\beta}_0$ and $\hat{\beta}_1$, then the origianl parameters can be estimated by $\hat{c} = \exp(\hat{\beta}_0)$ and $\hat{r} = \hat{\beta}_1 + 1$.

3. If we have a 95% confidence interval for β_0 given by $\beta_0 \in (-2.31, 4.16)$, the corresponding confidence interval for c would be $c \in \left(\text{e}^{-2.31}, \text{e}^{4.16}\right) = (0.0993, 64.072)$.

```
exp(c(-2.31, 4.16))
```

4. To test whether or not a city has *either* supra-linear scaling *or* sub-linear scaling, we want to determine whether or not $r = 1$. To do this, we could use the information from our linear model and test $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$. If there is significant evidence to reject H_0, then we would conclude that the city does not have linear scaling.

5. Using $n = 12$ observations, we have $\hat{\beta}_1 = 0.36$ and $\text{se}(\hat{\beta}_1) = 0.2$, and so our corresponding t statistic is $\mathcal{T} = \hat{\beta}_1/\text{se}(\hat{\beta}_1) = 0.36/0.2 = 1.8$. The critical value of this t test is $k = T_{10}^{-1}(0.975) = 2.228$. Because $\mathcal{T} < k$ and $\mathcal{T} > -k$, we fail to reject H_0. The data indicates that this city has linear scaling.

Question 5

Question 6

Suppose we have the linear relationship $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\epsilon \sim \text{N}(\mathbf{0}, \sigma^2\mathbf{I})$, and there are a maximum of k variables that can be used from $\mathbf{X}$. We would like to infer how many variables to include in our model. Let $f_p(\mathbf{X})$ denote the linear model that is fit using p predictors.

1. For a given number of predictors p, Mallows C_p is given by

$$C_p = \frac{\|\mathbf{y} - f_p(\mathbf{X})\|^2}{\hat{\sigma^2}} + 2p - n = (n-k)\frac{\|\mathbf{y} - f_p(\mathbf{X})\|^2}{\|\mathbf{y} - f_k(\mathbf{X})\|^2} + 2p - n.$$

That is, the value of $\hat{\sigma}^2$ is the mean-squared error for the model fit with all k predictors. Intuitively, for different values of p we will get a different value of C_p; the value of $\|\mathbf{y} - f_p(\mathbf{X})\|^2$ will change for each value of p, and we also have C_p increase by 2 for each additional variable considered. We choose the value of p that gives us the lowest C_p, and then use some kind of selection criteria (e.g. stepwise selection or the lasso) to determine *which* of the p predictors we should use.

2. The idea behind using AIC and BIC for model selection is the same as Mallows C_p, with the exeption of using a different formula. For AIC and BIC, we need to assume that the errors are normally distributed (as we have). For $p = 1, \ldots, k$, we find the value of AIC(p) and BIC(p) and choose the p that gives us the lowest criteria.

3. While similar in concept, AIC and BIC have fundamentally different derivations, which result in different formulas to use for the criteria and different tendencies for model selection. The formulas for each are given by

$$\text{AIC}(p) = 2p - 2\max(\ell; p) \quad \text{and} \quad \text{BIC}(p) = p\log n - 2\max(\ell; p);$$

here $\max(\ell; p)$ is the maximum possible value of the log-likelihood for the model with p predictors There are many more sophisticated justifications to using one or the other (that I was unable to wrap my head around), but practically, AIC is better to use if you are interested in model *prediction*, while BIC is better to use if you are interested in model *inference*. Also, BIC tends to choose smaller models than AIC or Mallows C_p.

4. Consider the new selection criteria $\text{D} = 2\|\hat{\boldsymbol{\beta}}\|_1 - 2\max(\ell; p)$. Compared to AIC, we see that both criteria give the same weight to $\max(\ell; p)$. The difference lies in how the coefficients play a role; by replacing p with $\|\hat{\boldsymbol{\beta}}\|_1$, we are changing the nature of the critetia. Whereas AIC would increase as the *number of parameters* increased, D will increase as the *total magnitute of each of the parameters* increases.

5. The nature of the criteria D suggests that it is much more similar to the lasso, which chooses the model parameters by penalizing large estimated coefficients. Indeed, the loss function for the lasso with p predictors is

$$P(\boldsymbol{\beta}; \lambda) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1,$$

which also incorperates the 1-norm.