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. 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 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.

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