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 $\hat{\sigma}^2 = (7930/364)^2 = 474.6173$.

```
(7930 / 364)^2
```

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 $(= \log_{10} X)$.