STAT 504: Applied Regression
Problem Set 1

Instructions: Submit your answers in a *single pdf file*. Your submission should be readable and well formatted. **Handwritten answers will not be accepted**. You can discuss the homework with your peers, but *you should write your own answers and code.* ***No late submissions will be accepted.*** The problem sets are long, so please plan accordingly.

Notation

In problem sets, we will maintain the notation used in class, in which we denote both the probability density function (PDF) or the probability mass function (PMF) of a random variable X by $p(x)$:

$$p(x) := P(X = x) \qquad \text{or} \qquad p(x) := \frac{dF(X)}{dX}\Big|_{X=x}$$

1 R/Python computing (10 points)

This exercise is just to make sure you have a working environment in `R` or `Python` early on in the class. You should be able to: (i) perform a simple simulation; (ii) fit a linear regression model; and (iii) draw a scatter plot of the data with the regression line.

Either in `R` or `Python` do the following:

(a) Simulate 100 draws from $X \sim N(0, 1)$, $\epsilon \sim N(0, 1)$, and $Y = 10 + 5 \times X + \epsilon$.

(b) Fit a linear regression model (ordinary least squares) regressing Y on X.

(c) Make a scatter plot with X in the horizontal axis and Y in the vertical axis. Draw the regression line in the scatter plot.

You should provide both your code and the output in your answer.

2 Probability spaces (10 points)

These are review questions of basic probability theory.

(a) Consider an experiment in which we roll a fair six-sided die. Define the sample space Ω, the event space S and the probability measure $P : S \to \mathbb{R}$ of the experiment.

(b) Suppose a researcher randomly picks one individual out of a population of 1000 people. In this population, 200 are Republicans, 400 are Democrats, and the remainder are Independents. After picking someone at random, the researcher records the Party ID of the person. Define the sample space Ω, the event space S and the probability measure $P : S \to \mathbb{R}$ of this random generative process.

3 Univariate random variables (30 points)

These are review questions of basic probability theory. For all questions bellow (except (e)), consider a continuous random variable X and scalars a and b.

3.1 (5 points)

(a) How is $\mathbb{E}[X]$ defined? (expected value)

(b) How is $\text{Var}[X]$ defined? (variance)

(c) Show that $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

(d) How is $\text{SD}[X]$ defined? (standard deviation)

(e) Show that $\mathbb{E}[g(X)] = \sum_x g(x)p(x)$. (Law of the Unconscious Statistician)

(f) Show that $\mathbb{E}[a + bX] = a + b\,\mathbb{E}[X]$.

(g) Show that $\text{Var}[a + bX] = b^2 \text{Var}[X]$.

(h) Show that $\text{SD}[a + bX] = |b|\,\text{SD}[X]$.

3.2 (5 points)

Prove Markov's and Chebyshev's inequality. Explain in words what Chebyshev's inequality mean.

3.3 (20 points)

Consider the hypothetical probability of snowfall in Seattle of Table 1. This data is provided in the file `snow.csv`. Let X denote the random variable "inches of snow." For all the questions below, do not compute "by hand", use `R` or `Python` to perform the computation.

(a) Draw both the PMF and the CDF of X.

(b) Compute $\mathbb{E}[X]$, $\text{Median}[X]$, $\text{Mode}[X]$, $\text{Var}[X]$. Compute the 95% percentile of X.

Snow (inches)	Prob
0	0.40
1	0.10
2	0.08
3	0.04
4	0.05
5	0.04
6	0.04
7	0.04
10	0.02
11	0.04
12	0.04
15	0.02
17	0.04
18	0.02
23	0.02
31	0.01

Table 1: Snowfall in Seattle.

(c) Compute the odds of snowing.

(d) Suppose you are asked to make a point prediction for the next snowfall. What are the "best" predictions you could make? How does that depend on your definition of "best"?

(e) If you use $\mathbb{E}[X]$ for making your prediction, what is the expected squared error?

(f) Construct a 95% prediction interval for the next snowfall.

4 Best predictors (univariate) (20 points)

Consider a random variable X and a predictor c for X.

(a) Show that the mean squared error of c can be decomposed as:

$$\mathbb{E}[(X-c)^2] = \text{Var}[X] + (\mathbb{E}[X]-c)^2$$

(b) Using the result above, explain why $\mathbb{E}[X]$ is the best predictor of X (in the MSE sense):

$$\mathbb{E}[X] = \arg\min_{c\in\mathbb{R}} \mathbb{E}[(X-c)^2]$$

(c) Here you may assume that X is a continuous random variable. Show that the median is the best predictor if we consider the mean absolute error. That is:

$$\text{Median}[X] = \arg\min_{c\in\mathbb{R}} \mathbb{E}[|X-c|]$$

(d) For a discrete random variable X, show that $\text{Mode}[X]$ minimizes the probability of making an error (or, maximizes the probability of a perfect prediction), that is:

$$\text{Mode}[X] = \arg\max_{c \in \mathbb{R}} P(X - c = 0)$$

5 Bivariate random variables (20 points)

5.1 (10 points)

For all questions bellow, consider continuous random variables X, Y, and Z, as well as scalars, a, b, c and d.

(a) Show that $\mathbb{E}[a + bX + cY] = a + b\,\mathbb{E}[X] + c\,\mathbb{E}[Y]$.

(b) What is the definition of $\mathbb{E}[Y \mid X = x]$? Explain what it means in plain English.

(c) What is the definition of $\text{Var}[Y \mid X = x]$? Explain what it means in plain English.

(d) What is the definition of $\text{Cov}(X, Y)$? (covariance)

(e) Show that $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$, and that $\text{Cov}(X, X) = \text{Var}(X)$.

(f) Show that $\text{Cov}(bX, cY) = bc\,\text{Cov}(X, Y)$

(g) Show that $\text{Var}(a + bX + cY) = b^2\,\text{Var}(X) + c^2\,\text{Var}(Y) + 2bc\,\text{Cov}(X, Y)$

(h) Show that $\text{Cov}(Y + X, Z) = \text{Cov}(Y, Z) + \text{Cov}(X, Z)$.

(i) What is the definition of $\text{Cor}(X, Y)$? (correlation)

(j) Show that $\text{Cor}(a + bX, c + dY) = \frac{bd}{|bd|}\,\text{Cor}(X, Y)$

5.2 (10 points)

Consider the hypothetical joint PMF of X (income in thousands of dollars) and Y (savings rate) given in Table 2. The data is provided in the file `income-savings.csv`. For all the questions below, do not compute "by hand", use `R` or `Python` to perform the computation.

(a) Explain in plain English what the joint distribution means.

(b) Compute the marginal distributions of X and Y.

(c) Compute the conditional distributions of Y given X, for every value of $X = x$.

(d) Compute the conditional expectation of Y given X, for every value of $X = x$ (the CEF). Show numerically that $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$.

	X									
Y	0.5	1.5	2.5	3.5	4.5	5.5	6.7	8.8	12.5	17.5
.50	.001	.011	.007	.006	.005	.005	.008	.009	.014	.004
.40	.001	.002	.006	.007	.010	.007	.008	.009	.008	.007
.25	.002	.006	.004	.007	.010	.011	.020	.019	.013	.006
.15	.002	.009	.009	.012	.016	.020	.042	.054	.024	.020
.05	.010	.023	.033	.031	.041	.029	.047	.039	.042	.007
0	.013	.013	.000	.002	.001	.000	.000	.000	.000	.000
$-.05$	.001	.012	.011	.005	.012	.016	.017	.014	.004	.003
$-.18$	.002	.008	.013	.006	.009	.008	.008	.008	.006	.002
$-.25$	.009	.009	.010	.006	.009	.007	.005	.003	.002	.003

Table 2: Joint distribution of X and Y.

(e) Compute the best linear predictor of Y given X (BLP).

(f) Draw a plot of the BLP and the CEF (together).

6 Election Forecasting (10 points)

For this exercise we will consider the dataset `hibbs.dat`, as appeared in ?. Here you will run your first regression model on real data. Our goal is to construct a simple model to forecast elections based solely on economic growth. The variable `vote` represents the incumbent party's vote percentage of the two-party vote; the variable `growth` represents the average personal income growth in the previous years. For now, do not worry about sampling uncertainty, we will revisit this example later in the course.

(a) Before looking at the data, do you expect to see an association between `vote` and `growth`? In what direction?

(b) Make a scatter plot of `vote` against `growth`. Fit a linear regression model to predict `vote` using `growth`. Draw the regression line in the scatter plot. Does the model seem to provide a good fit?

(c) Interpret the regression line and the estimated regression coefficients.

(d) In the 2016 presidential election of Hillary Clinton vs. Donald Trump, Hillary (the candidate of the incumbent party) received 51.1% of the popular vote (but lost the electoral college). At the time, the average growth was 2%. What is the prediction that our model produces in this case? Is it close to the true vote share?

Acknowledgements: Q2(a) based on A&M (PL500). Table 2 based on [Goldberger, 1991, Chapter 1].

References

Peter M Aronow and Benjamin T Miller. *Foundations of agnostic statistics.* Cambridge University Press, 2019.

Arthur Stanley Goldberger. *A course in econometrics.* Harvard University Press, 1991.