GRA 4153 – Exercises[1]

1 Probability

1.1 Probabilities, random variables

(1.1.1) A fair die is thrown until a 6 appears. What is the probability it must be thrown at least k times?

(1.1.2) For $x \in \mathbb{R}^K$, let $f_\theta(x) \coloneqq h(x)\exp(\eta(\theta)'T(x) - A(\theta))$ for functions h, η, T, A. When is this a probability density function?[2]

(1.1.3) For each $i = 1, \ldots, K$, let f_i be probability density functions (resp. probability mass functions) and define $f(x) \coloneqq \sum_{i=1}^K w_i f_i(x)$ where $w_i \geq 0$ and $\sum_{i=1}^K w_i = 1$. Show that f is a probability density (resp. probability mass) function.[3]

(1.1.4) Let X be a Poisson random variable with mass function $f(x) = \lambda^x \exp(-\lambda)/x!$, $x = 0, 1, \ldots$ for a $\lambda > 0$. Find the probability that X is odd.

(1.1.5) Prove that $F(x) \coloneqq (1 + \exp(-x))^{-1}$ $(x \in \mathbb{R})$ is a CDF.

(1.1.6) *Show that any CDF F, i.e. $F(x) \coloneqq P(X \leq x)$, can have at most a countable number of discontinuities.

1.2 Expectations

(1.2.1) Show that $\mathbb{E}\,\alpha = \alpha$ for any non-random α.

(1.2.2) Let X be the sum of two rolls of a fair die, as in Example 2.1. What is the mean and variance of X?

(1.2.3) X is uniformly distributed on $[a, b]$ if its density is $f(x) = \frac{1}{b-a}$. Compute the mean and variance of X.

(1.2.4) Calculate the mean of $X \sim t(\nu)$. Are restrictions on ν required for the mean to exist?

(1.2.5) Prove Proposition 2.6 for the discrete case

(1.2.6) Prove Lemma 2.1

(1.2.7) Prove Lemma 2.2

(1.2.8) Let X and Y be random variables with $\mathbb{E}\,|X| < \infty$, $\mathbb{E}\,|Y| < \infty$ and let $X \wedge Y \coloneqq \min\{X, Y\}$ and $X \vee Y \coloneqq \max\{X, Y\}$. Show that $\mathbb{E}(X \vee Y) = \mathbb{E}\,X + \mathbb{E}\,Y - \mathbb{E}(X \wedge Y)$. [Hint: What is $(X \vee Y) + (X \wedge Y)$?]

[1] If any typos are found, please let me know at adam.lee@bi.no

[2] I.e. when is $f(x) \geq 0$ and such that its integral is equal to one?

[3] I.e. show that $f(x) \geq 0$ and its integral / sum is equal to one in the density / mass case respectively.

1.3 Conditioning & independence

(1.3.1) If P is a probability and B an event with $P(B) > 0$ show that $P(\cdot|B)$ is also a probability.

(1.3.2) If $P(B \cap C) > 0$ show that $P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C)$.

(1.3.3) Prove Lemma 2.3

(1.3.4) Prove Lemma 2.4

(1.3.5) Prove Lemma 2.5

(1.3.6) Prove Corollary 2.1

(1.3.7) Prove Proposition 2.10

(1.3.8) Prove the "law of total variance": if $\mathrm{Var}(X) < \infty$ then $\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}(\mathbb{E}[X|Y])$ [Hint: $\mathrm{Var}(X|Y) = \mathbb{E}((X - \mathbb{E}[X|Y])^2|Y)$].

1.4 Key results

(1.4.1) Prove Corollary 2.2

(1.4.2) *Let X be a random vector. Prove that its characteristic function exists.

(1.4.3) *Complete the proof of Proposition 2.13

(1.4.4) *Let ψ be a characteristic function. Prove that ψ is uniformly continuous on $\mathbb{R}$. [Hint: use Proposition 2.13]

(1.4.5) *Given that (i) if $Z \sim \mathcal{N}(\mu, \Sigma)$, $Z = AX + \mu$ for $X \sim \mathcal{N}(0, I)$ and $\Sigma = AA'$ and (ii) the characteristic function of a standard normal random variable is $\exp(-t^2/2)$, compute the characteristic function of $X \sim \mathcal{N}(\mu, \Sigma)$.

1.5 Stochastic convergence

(1.5.1) *Prove Lemma 2.7

(1.5.2) *Fill in the missing details in the proof of Theorem 2.8

(1.5.3) Complete the proof of Theorem 2.9

(1.5.4) *Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}$. Prove: $x_n \to x$ if and only if each subsequence $(x_{n_m})_{m\in\mathbb{N}}$ has a further subsequence which converges to x.[4]

(1.5.5) Prove Lemma 2.8

(1.5.6) Show: (i) if $Z_n \rightsquigarrow Z$ then $Z_n = O_P(1)$; (ii) if $Z_n = O_P(1)$ then $o(\|Z_n\|) = o_P(1)$ and $o_P(\|Z_n\|) = o_P(1)$.

[4]This is true much more generally: if $(x_n)_{n\in\mathbb{N}}$ is a sequence in a topological space this remains true.

(1.5.7) Show that X_n defined in Example 2.13 satisfies $X_n \xrightarrow{P} 0$.

(1.5.8) Show that if $(X_n)_{n\in\mathbb{N}}$ is uniformly integrable then $\sup_{n\in\mathbb{N}} \mathbb{E}\,\|X_n\| < \infty$.

(1.5.9) *Fill in the details for the general case of Theorem 2.14. Hint: a random variable X may be written as $X = Y + Z$ with $Y = X\mathbf{1}\{X \geq 0\}$ and $Z = X\mathbf{1}\{X < 0\}$.

(1.5.10) Prove Lemma 2.10

(1.5.11) Prove Lemma 2.11

(1.5.12) Prove Lemma 2.12

(1.5.13) Prove Lemma 2.13

(1.5.14) Prove Lemma 2.14

(1.5.15) Prove Lemma 2.16

(1.5.16) Fill in the missing details in the proof of Proposition 2.14

(1.5.17) Prove Corollary 2.5

(1.5.18) Prove the following WLLN:

THEOREM [L_2 WLLN]: *If $(X_n)_{n\in\mathbb{N}}$ are uncorrelated random variables with $\mathbb{E}\,X_i = \mu$ and $\mathrm{Var}(X_i) \leq C < \infty$ then $\frac{1}{n}\sum_{i=1}^n X_i$ converges to μ in L_2 and in probability.*

(1.5.19) Draw $M = 5000$ samples of size $n = 400k$ for $k = 1, \ldots, 7$ from the model:

$$X = \mu + \epsilon, \qquad \epsilon \sim t(15), \quad \mu = 1.$$

Let $\hat{\mu}_n \coloneqq \frac{1}{n}\sum_{i=1}^n X_i$ and calculate (i) the empirical probability that the distance between $\hat{\mu}_n$ and μ is greater than 0.05: $\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{|\hat{\mu}_{n,i} - \mu| > 0.05\}$ and (ii) the (empirical) mean squared error $\frac{1}{n}\sum_{i=1}^n (\hat{\mu}_{n,i} - \mu)^2$ in our sample. Plot these and comment on how the results relate to Theorem 2.15 and the preceding exercise

(1.5.20) Prove Corollary 2.6

(1.5.21) *Prove Corollary 2.7

(1.5.22) *Prove Lemma 2.17

2 Statistics

2.1 Statistics, Sufficiency, Completeness & Ancilliarity

(2.1.1) In the context of Example 3.1 which of the following are statistics: (a) $\frac{1}{n}\sum_{i=1}^n X_i$, (b) $\sum_{i=1}^n X_i$, (c) $\frac{1}{n}\sum_{i=1}^n X_i^2$, (d) $\frac{1}{n}\sum_{i=1}^n (X_i - \theta)$, (e) $\frac{1}{n}\sum_{i=1}^n (X_i^2 - 1)$?

(2.1.2) In the context of Example 3.1, find a minimal sufficient statistic.

(2.1.3) Let $X \sim \text{Poisson}(\theta)$ and $Y \sim \text{Poisson}(\theta^2)$, independent, with $\theta \in (0, \infty)$. (a) Find a minimal sufficient statistic for the family of joint distributions for $Z = (X, Y)$. (b) Is this minimal sufficient statistic complete?

(2.1.4) *Let $Z = (Z_1, \ldots, Z_n)$ be i.i.d. standard normal random variables. Show that $\|Z\|$ and $Z/\|Z\|$ are independent. [Hint: consider an appropriate model where each Z_i has a normal distribution depending on some θ and use Basu's Theorem].

(2.1.5) Consider the family $\{f_p : p \in [0, 1]\}$ where f_p is the Bernoulli pmf of Example 2.2. Show this is an exponential family and write it in canonical form.

(2.1.6) Consider the family $\{p_\lambda : \lambda \in (0, \infty)\}$ where p_λ is the Poisson pmf of Example 2.4. Show this is an exponential family and write it in canonical form.

2.2 Statistical inference and performance criteria

(2.2.1) Prove the Bias-Variance decomposition in Example 3.10 [Hint: The general case follows from the univariate case which can be shown by applying $\text{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2$ to $Z = T(X) - g(\theta)$]

(2.2.2) Let $X = (X_1, \ldots, X_n)$ and consider the family $\{p_\lambda^n : \lambda \in (0, \infty)\}$ where p_λ is the Poisson pmf of Example 2.4. Consider the estimator $T(X) = X_1$. Show this is unbiased. Use the Rao-Blackwell Theorem to find a better estimator. [* Derive an explicit expression for the estimator.]

(2.2.3) In the setting of Example 3.12, show that $a = (n+2)/(n+1)$ minimises $R(\theta, \delta_a)$. Calculate the bias of the corresponding estimator δ_a.

(2.2.4) In the setting of Example 3.13 show that conditional on $T = t$, $X_1 = t$ with probability $1/n$ and X_1 is uniformly distributed on $(0, t)$ with probability $(n-1)/n$.

(2.2.5) Consider a random sample $X = (X_1, \ldots, X_n)$ where each $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Calculate the information matrix $I(\theta)$ where $\theta = (\mu, \sigma^2)$ and obtain UMVU estimators for μ and σ^2 respectively.

(2.2.6) Suppose p_θ is a pdf for each θ, that (i) of Theorem 3.7 holds and that p_θ can be twice differentiated under the integral sign in that

$$\nabla_\theta \nabla_{\theta'} \int p_\theta(x)\,\mathrm{d}x = \int \nabla_\theta \nabla_{\theta'} p_\theta(x)\,\mathrm{d}x.$$

Show that $-I(\theta) = \mathbb{E}_{P_\theta}[\nabla_{\theta'} \ell_\theta]$. This is called the *information equality*.

(2.2.7) Suppose that $S(X)$ is a confidence set for $g(\theta)$ such that $P_\theta(g(\theta) \in S(X)) = 1 - \alpha$ for all $\theta \in \Theta$. Form a test, φ, as in (21). What is the size of φ?

(2.2.8) Show: k_α is the 1-α quantile of the $\chi^2(1)$ distribution if and only if $k_\alpha^{1/2}$ is the 1-$\alpha/2$ quantile of the $\mathcal{N}(0,1)$ distribution.

(2.2.9) Verify the form of $S(X)$ in Example 3.15.

2.3 Asymptotic statistics

(2.3.1) Prove Proposition 3.4 [Hint: Example 3.10 may be useful]

(2.3.2) Suppose that $X_1, X_2, \ldots$ are i.i.d. Poisson(λ) random variables. Show that $\hat{\lambda}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is consistent for λ.

(2.3.3) Draw some large integer n samples from (i) a standard Normal distribution and (ii) from a Cauchy distribution. Plot the cumulative sample average over $1, \ldots, n$ (i.e. plot $X_1, (X_1 + X_2)/2, (X_1+X_2+X_3)/3, \ldots, n^{-1}\sum_{i=1}^n X_i$). Describe the observed behaviour and whether the sequence appears to follow the WLLN. If not, why not?

(2.3.4) In example 3.18 show that $s_n^2 \xrightarrow{P} \sigma^2$.

(2.3.5) Compare the (asymptotic) t-test defined in (27) to the test developed in Example 3.14.

(2.3.6) *Fill in the missing details in the proof of Proposition 3.5.

(2.3.7) Compare the (asymptotic) confidence interval S_n in Example 3.22 with that developed in Example 3.15.

(2.3.8) Suppose that $\hat{\theta}_n$ as in (31) exists. Show that the two definitions in (31) and (32) are equivalent.

(2.3.9) Let $\{P_\theta : \theta \in \Theta\}$ be a model in which θ is not identified. That is, there are $\theta_1 \neq \theta_2$ with $P_{\theta_1} = P_{\theta_2}$. Show that no estimator can be consistent in this model.

(2.3.10) Show that if $X \sim$ Poisson(λ) then $\mathbb{E}\, X = \mathrm{Var} X = \lambda$.

(2.3.11) Show that if $\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}(0, V)$ (under P_θ) then $\hat{\theta}_n \xrightarrow{P_\theta} \theta$.

(2.3.12) Suppose that $X_1, X_2, \ldots$ are i.i.d. with $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Let $\theta = (\mu, \sigma^2)$. Find the MLE of θ and show directly (i.e. without using Theorems 3.8, 3.9) that it is asymptotically normal.

(2.3.13) Suppose that $X_1, X_2, \ldots$ are i.i.d. with logistic pdf

$$p_\theta(x) = \frac{\exp(-(x-\theta))}{(1+\exp(-(x-\theta)))^2}, \qquad \theta \in (-\infty, \infty).$$

Write a function in Python which computes the log-likelihood $l(\theta)$. Write a function in Python which finds the MLE of θ. [Hint: you might find the Scipy.optimize library useful] Perform a Monte-Carlo simulation exercise as follows: draw M data sets of n samples $X_1, \ldots, X_n$ which are logistically distributed [Hint: numpy.random.logistic]. For

each sample calcuate the MLE $\hat{\theta}_n$ and the distance of $\hat{\theta}_n$ from θ. Repeat this for various sample sizes n and report your findings (e.g. graphically).

(2.3.14) Prove Lemma 3.2.

3 Linear regression

3.1 Least squares estimation

(3.1.1) In the proof of Proposition 4.1 verify that (i) $X'X$ is of full rank, (ii) C is a linear subspace of $\mathbb{R}^n$ and (iii) $(y - X\hat{\beta})'X\beta = 0$ for all $\beta \in \mathbb{R}^K$ implies that $(y - X\hat{\beta})'X = 0$.

(3.1.2) Provide an alternative proof of Proposition 4.1 by minimising $S(\beta)$ using the first- and second- order (derivative) conditions.

(3.1.3) Provide a proof of Lemma 4.1.

(3.1.4) In the setting of Theorem 4.1, prove that $M_{X_1}M_X = M_X$, $M_X M_{X_1} X_2 = 0$ and that $M_{X_1}X_2$ has full column rank.

(3.1.5) Consider the regression

$$y = X_1\alpha + X_2\beta + \epsilon,$$

where $X_1 = \mathbf{1}_n$ is a n-vector of 1s. Show that the least squares estimate of β is identical to that obtained from the regression

$$\tilde{y} = \tilde{X}_2\beta + u,$$

where $\tilde{X}_2 = X_2 - \mathbf{1}_n \left[\frac{1}{n}\sum_{i=1}^n X_{2,i}\right]$ and $\tilde{y} = y - \mathbf{1}_n \left[\frac{1}{n}\sum_{i=1}^n y_i\right]$.

(3.1.6) Write a function which accepts a n dimensional vector y and a $n \times K$ matrix X and returns the OLS estimate $\hat{\beta}$, using (34).

(3.1.7) Write a function which accepts a n dimensional vector y and a $n \times K$ matrix X and returns the OLS estimate $\hat{\beta}$, by numerically minimising (35) [hint: `scipy.optimize.minimize` can be used to minimise $S(\beta)$]. Verify the estimate matches that computed based on the explicit solution (34) (up to numerical precision). Verify that both estimates match that obtained using `statsmodels.OLS`.

3.2 Statistical properties of the least squares estimator

(3.2.1) Under Assumptions 4.1 - 4.4 show that $\text{Cov}\left(\hat{\beta}, \hat{\epsilon}\middle| X\right) = 0$, where $\hat{\epsilon} \coloneqq y - X\hat{\beta}$.

(3.2.2) Let D be a $m \times k$ matrix. Show that DD' is positive semi-definite.

(3.2.3) Use Theorem 4.4 to show that under the same conditions, for any linear unbiased estimator $\tilde{\beta}$, $\text{Var}\left[\hat{\beta}\right] \leq \text{Var}\left[\tilde{\beta}\right]$.

(3.2.4) Fix some X matrix of full column rank, β vector and repeatedly draw M samples of y satisfying Assumptions 4.1, 4.2 and 4.4, for some (large) integer M. Calculate the mean (over the M samples) and variance of $\hat{\beta}$ and compare to the results of Propositions 4.2 and 4.3.

(3.2.5) Suppose that Assumptions 4.1, 4.2, 4.3 and 4.5 hold. Show that

$$\mathrm{Var}\left[\hat{\beta}\middle|X\right] = \sigma^2(X'X)^{-1}X'V(X)X(X'X)^{-1}.$$

(3.2.6) Let $\tilde{S}(\beta) \coloneqq (y - X\beta)'V(X)^{-1}(y - X\beta)$. Show that $\tilde{\beta}$ [as in (46)] minimises $\tilde{S}(\beta)$.

(3.2.7) Prove Proposition 4.5.

(3.2.8) Fix some X matrix of full column rank, β vector and repeatedly draw M samples of y satisfying Assumptions 4.1, 4.2 and 4.5 (but not 4.4), for some (large) integer M. Calculate the variance (over the M samples) of $\hat{\beta}$ and $\tilde{\beta}$ and compare to the result of Theorem 4.5.

(3.2.9) Draw some large integer n samples from (i) a standard Normal distribution and (ii) from a Cauchy distribution. Plot the cumulative sample average over $1, \ldots, n$ (i.e. plot $X_1, (X_1 + X_2)/2, (X_1+X_2+X_3)/3, \ldots, n^{-1}\sum_{i=1}^n X_i$). Describe the observed behaviour and whether the sequence appears to follow the WLLN. If not, why not?

(3.2.10) Suppose Assumptions 4.6, 4.7, 4.8 and 4.9 hold and there are two regressors (so $X_i = (x_{1,i}, x_{2,i})' \in \mathbb{R}^2$). Suppose that $\check{\beta}_{1,n}$ is the estimator formed by regressing y on x_1 only:

$$\check{\beta}_{1,n} \coloneqq \left(\sum_{i=1}^n x_{1,i}^2\right)^{-1}\left(\sum_{i=1}^n x_{1,i}y_i\right).$$

Show that this converges in probability to a number α and give a formula for α in terms of β_1, β_2 and $\Sigma \coloneqq \mathbb{E}[X_iX_i']$.

(3.2.11) Draw M data sets (y, X) of n samples which satisfy Assumptions 4.6, 4.7, 4.8 and 4.9. For each of the $m = 1, \ldots, M$ data sets calculate the average difference of $\hat{\beta}_n$ from β. Calculate the empirical probability that the difference exceeds a small value $\varepsilon > 0$ (try a few different values for ε). How does this change with n? Discuss with reference to Proposition 4.6.

(3.2.12) Draw M data sets (y, X) with $K = 1$ of n samples which satisfy Assumptions 4.6, 4.7, 4.8, 4.9 and 4.10. For each of the $m = 1, \ldots, M$ data sets calculate $\hat{\beta}_n$. Plot a histogram of the M (i) $\hat{\beta}_n$ estimates, (ii) $(\hat{\beta}_n - \beta)$ and (iii) $\sqrt{n}(\hat{\beta}_n - \beta)$. How do these histograms change with n? Discuss with reference to Proposition 4.7.

(3.2.13) Prove Proposition 4.8.

(3.2.14) Prove Proposition 4.9.

(3.2.15) Let $\tilde{\sigma}_n^2$ be the estimator defined in (43) and let $\hat{\sigma}_n^2$ be the estimator defined in (56) where $\check{\beta}_n = \hat{\beta}_n$ (i.e. the OLS estimator). Show that in the setting of Lemma 4.4, $\tilde{\sigma}_n^2 \xrightarrow{P} \sigma^2$. Show that in the setting of Proposition 4.4, $\mathbb{E}\left[\hat{\sigma}_n^2 | X\right] \neq \sigma^2$.

(3.2.16) *Where are Assumptions 4.8 and 4.12 used in the proof of Lemma 4.5? Show that the bounds in equation (61) hold.

(3.2.17) Suppose that $0 < \mathbb{E}[\epsilon_i^2 | X_i] = \sigma^2 < \infty$ and Assumptions 4.6, 4.7, 4.8, 4.9 and 4.10 hold.

(a) Prove that $M^{-1}\Sigma M^{-1} = \sigma^2 M^{-1}$.

(b) Let $\hat{\sigma}_n^2$ be defined as (43), with any consistent estimator $\check{\beta}_n$ in place of $\hat{\beta}$. Prove that

$$\hat{\sigma}_n^2 M_n^{-1} \xrightarrow{P} \sigma^2 M^{-1}.$$

(3.2.18) Formulate the Wald statistic for a test of the hypothesis $\beta_1 + \cdots + \beta_k = 0$.

(3.2.19) Formulate the Wald statistic for a test of the hypothesis $\|\beta\| = 0$.

(3.2.20) Draw M data sets (y, X) with $K = 1$ of n samples which satisfy Assumptions 4.6, 4.7, 4.8, 4.9 and 4.10, with $\beta_1 = 0$. For each data set conduct a Wald test of the hypothesis $\beta_1 = 0$, at $\alpha = 0.05$. Compute average rejection rate of the test and compare it to α.

(3.2.21) Repeat the previous exercise B times, each time drawing the data the same way with one difference: for each $b = 1, \ldots, B$ choose a different β_1 when drawing the data such that $\beta_{1,1}, \ldots, \beta_{1,B}$ are equally spaced in $[-b, b]$ (and one $\beta_{1,b} = 0$) [in each case the hypothesis remains $\beta_1 = 0$]. Plot the average rejection rate in each run of the exercise against $\beta_{1,b}$, to see the finite-sample power against these alternatives.

(3.2.22) Import the Star98 dataset using the code:

```
import statsmodels.api as sm
df = sm.datasets.star98.load_pandas().data
```

(a) Regress "NABOVE" on (a constant,) "LOWINC", "AVYRSEXP", "PERSPENK" and "PTRATIO" using `statsmodels.OLS`. (b) Interpret the results (variable descriptions can be found here). Test whether (c) the coefficient on "LOWINC" differs from 0 at the 5% level using a wald test; (d) whether all four coefficients are different from 0 at the 5% level using a Wald test.[5]

3.3 Evaluation of estimators and predictions

(3.3.1) Verify the last equality in the display in the proof of Proposition 4.10.

(3.3.2) Proposition 4.10 is given for the case where $y \in \mathbb{R}$. Show that this suffices for the general case (i.e. write an analogous statement for the case where $y \in \mathbb{R}^K$ with $K \in \mathbb{N}$ arbitrary and use the result of Proposition 4.10 to prove this statement).

[5] If `res` is the result of a call to `sm.OLS(y, X).fit()` the method `res.wald_test` performs a Wald test.

(3.3.3) Show that a sequence $(X_n)_{n\in\mathbb{N}}$ of i.i.d. random variables with $\mathbb{E}\,|X_1| < \infty$ is uniformly integrable.

(3.3.4) State and prove a version of Proposition 4.11 for the case where $n > 1$.

3.4 Regularisation

(3.4.1) Prove that for any symmetric positive-semidefinite matrix M and any $\lambda > 0$, $M + \lambda I$ is positive definite.

(3.4.2) Show that the ridge regression estimator is numerically equal to the OLS estimator calculated on an augmented data set, where we add K zeros to the end of the y vector and K additional rows $\sqrt{\lambda}I_K$ to the end of the X matrix.

(3.4.3) Show that for $\lambda > 0$, $I_K - \lambda(X'X + \lambda I_K)^{-1} = (X'X + \lambda I_K)^{-1}X'X$.

(3.4.4) If the covariates are orthogonal, i.e. $X'X = I$, the ridge regression estimator is a multiple of the OLS estimator: $\check{\beta}_\lambda = c\hat{\beta}$. Find c.

(3.4.5) Assume that $\mathbb{E}[\epsilon\epsilon'|X] = \sigma^2 I$ and show that the square of the norm of the (conditional) bias and the trace of the variance of the ridge regression estimator are increasing and decreasing (respectively) in λ.

(3.4.6) Lemma 4.6 demonstrates that the ridge regression estimator is biased (unless $\beta = 0$ or $\lambda = 0$). Supposing we had not calculated the conditional expectation, how can you reach the conclusion that the ridge estimator is biased based on its conditional variance?

(3.4.7) In the context of footnote 76, verify that $\frac{(x'b)^2}{x'Axb'A^{-1}b}$ attains its maximum at $x = \pm A^{-1}b$.

(3.4.8) If A, B are positive semidefinite matrices, then $\operatorname{tr}(AB) \geq 0$. If, in addition, A is positive definite and $B \neq 0$, $\operatorname{tr}(AB) > 0$.

(3.4.9) Suppose that Assumptions 4.6, 4.7, 4.8 and 4.9 hold. Is the ridge estimator consistent? What happens if $\lambda = \lambda_n = cn$ for some constant c? What about if $\lambda = \lambda_n = cn^{1/2}$?

(3.4.10) Suppose that Assumptions 4.6, 4.7, 4.8, 4.9 and 4.10 hold. What is the asymptotic distribution of $\sqrt{n}(\check{\beta}_\lambda - \beta)$?

(3.4.11) Using the scikit-learn library import the diabetes dataset:

```
from sklearn import datasets
X, y = datasets.load_diabetes(return_X_y=True)
```

Compare the OLS, Ridge and Lasso estimates (include all variables in X in the regresion) of the coefficients, using k-fold cross validation to choose the regularisation parameter for Ridge and Lasso.[6]

[6] E.g. using `LinearRegression`, `LassoCV` and `RidgeCV` from `sklearn.linear_model`.

3.5 Generalised linear models

(3.5.1) Prove Lemma 4.8.

(3.5.2) Verify the form of the log-likelihood in Example 4.16. Using $g^{-1}(z) = [1 + \exp(-z)]^{-1}$, provide an explicit expression for the log likelihood (i.e. substitute in for $p_i(X_i'\beta)$ and simplify). Show that g is the canonical link function.

(3.5.3) A commonly used alternative to the logit model described in Example 4.16 is the "probit" model. This uses the link function $g(p) = \Phi^{-1}(p)$ where Φ is the CDF of a standard normal random variable. Otherwise the model is the same as the logit model. Derive the log-likelihood (conditional on X_i) for the Probit model. Show that g is *not* the canonical link function.

(3.5.4) Verify the form of the log-likelihood in Example 4.17. Why may we equivalently maximise $\tilde{l}(\beta) = \sum_{i=1}^{n} [y_i X_i'\beta - \exp(X_i'\beta)]$? Show that g is the canonical link.

(3.5.5) In a GLM $g(\mu(X_i)) = \theta(X_i) = X_i'\beta$, i.e. $\mu(X_i) = \mathbb{E}[y_i|X_i] = g^{-1}(X_i'\beta)$.[7] Suppose that $f = g^{-1}$ is differentiable. What is the effect of a one unit change in $X_{i,k}$ on $\mathbb{E}[y_i|X_i]$, if everything else remains fixed?

(3.5.6) Pick one of the GLMs used as examples in the notes and run a simulation study to examine the asymptotic properties of (a) the MLE of the parameter β and (b) a test that some function $f(\beta) = 0$.

(3.5.7) Estimate the model in Example 4.18 using (i) Logit (as done in the notes), (ii) Probit and (iii) a linear regression model. In each case calculate the implied (average) effect of an increase of two years of education on the probability of being employed.

(3.5.8) Estimate the model in Example 4.19 using *months* instead of log(*months*). What do you find? Find an expression for the effect of one additional month at sea on the expected number of accidents in each of these two models.

4 Time Series

4.1 Stationarity

(4.1.1) Show the variance of a weakly stationary process is constant.

(4.1.2) Prove the statements in Example 5.1.

(4.1.3) Prove Lemma 5.1.

(4.1.4) Finish the argument in Example 5.3.

[7] An injective function always has a left inverse.

4.2 Fundamental processes

(4.2.1) Consider the process:

$$X_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}, \qquad \epsilon_t \sim \mathrm{WN}(0, \sigma^2).$$

Show that this is a stationary process with $\mathbb{E}\, X_t = \mu$ and the same autocovariance function as the MA(q) process in Lemma 5.2

(4.2.2) Consider the process

$$X_t = c + \phi X_{t-1} + \epsilon_t, \qquad \epsilon_t \sim \mathrm{WN}(0, \sigma^2),$$

with $|\phi| < 1$. Show that this has a stationary solution, and derive its mean and autocovariance function.

(4.2.3) Compute the autocovariance and autocorrelation function of a MA(2) process. Under what conditions on θ_1, θ_2 will there exist an invertible solution to the MA(2) equations?

(4.2.4) Compute the autocovariance and autocorrelation function of a stationary AR(2) process. Under what conditions on ϕ_1, ϕ_2 will will there exist a causal and stationary solution to the AR(2) equations?

(4.2.5) Which of the following processes are causal and/or invertible? (ϵ_t is white noise.)

(a) $X_t = 0.1X_{t-1} + \epsilon_t + 0.9\epsilon_{t-1}$

(b) $X_t = 0.2X_{t-1} + 0.5X_{t-2} + \epsilon_t$

(c) $X_t + 1.6X_{t-1} = \epsilon_t - 0.4\epsilon_{t-1} + 0.08\epsilon_{t-2}$

(4.2.6) For each of the causal time series in the previous exercise plot a realisation of $T = 400$ points (draw, say, 800 points initially and plot only the last 400; this is called "burn in") and their autocovariance function.

(4.2.7) Show that the random walk process $Y_t = \sum_{j=1}^{t} \epsilon_j$, $\epsilon_j \sim \mathrm{WN}(0, \sigma^2)$ is not stationary.

(4.2.8) Let $X_t = \phi X_{t-1} + \epsilon_t$ with $|\phi| > 1$. As seen in class, the unique stationary solution to this equation is given by $X_t = -\sum_{j=1}^{\infty} \phi^{-j} \epsilon_{t+j}$. Define $u_t \coloneqq X_t - \phi^{-1} X_{t-1}$. Show that (a) $u_t \sim \mathrm{WN}(0, \sigma^2)$, (b) there is a stationary, causal solution to the equation $X_t = \phi^{-1} X_{t-1} + u_t$.

4.3 Forecasting

(4.3.1) Verify that $\mathbb{E}\left[(X - \lambda' W) W'\right](\lambda - \alpha) = 0$ in the proof of Proposition 5.4.

(4.3.2) Perform the re-arrangement to reach the expression for λ as in Remark 5.1. Discuss the relationship with OLS.

(4.3.3) Prove (i) and (iii) in Proposition 5.5.

(4.3.4) Suppose that $(X_t)_{t\in\mathbb{Z}}$ is an AR(p) process. What is the best linear prediction of X_{t+h} given $X_{t-1}, \dots, X_1$ for $h \geq 1$?

(4.3.5) Suppose that $(X_t)_{t\in\mathbb{Z}}$ is an AR(p) process with intercept $c \neq 0$. What is the best linear prediction of X_{t+h} given $X_{t-1}, \dots, X_1, 1$ for $h \geq 1$?

(4.3.6) Suppose that $(X_t)_{t\in\mathbb{Z}}$ is an MA(1) process. Compute the best linear prediction $\mathbb{E}[X_{t+1}|X^t]$ using the innovations algorithm and the Durbin – Levinson algorithm.

(4.3.7) Implement both the Durbin – Levinson and innovations algorithm in Python.

(4.3.8) Draw artificial data from an ARMA(2, 3) model and predict h steps ahead for $h = 1, \dots, 4$ using the innovations algorithm.

4.4 Estimation and ARMA order selection

(4.4.1) Calculate $\Sigma(\beta)$ for a (causal, invertible) AR(1), MA(2) and ARMA(1, 1)

(4.4.2) Use your implementation of the innovations algorithm to implement the likelihood function $L_{X^n}(\gamma)$. Compare to a direct implementation of the form involving the determinant and inverse of Γ_n

(4.4.3) Draw data from an ARMA model and estimate ϕ, θ, σ^2 using a software package.[8]

(4.4.4) Draw data from an ARMA model and forecast $h = 1, \dots, 4$ periods ahead using a built-in implementation in a software package.[9]

(4.4.5) Draw data from an AR model, an MA models and a ARMA model and use the ACF, PACF and information criteria to select the model order.[10]

(4.4.6) Download the GDP (% change) and inflation series in the intro to time series notebook on the course Github. Estimate an ARMA(p, q) model for each of these series, using an appropriate method to choose p, q (each of which may be 0). Use your model to forecast GDP growth and inflation for the next 6 months.

[8] E.g. use `sm.tsa.arima.ARIMA(X, order = (p, 0, q)).fit()` from statsmodels.
[9] E.g. you might consider the `forecast` method of a ARIMA fit in statsmodels.
[10] See e.g. the function `sm.tsa.stattools.arma_order_select_ic` in statsmodels.