

STA 35C Statistical Data Science III

Midterm exam 1 solution

Instructor: Dogyoon Song

Problem 1 (40 points). Basic concepts

(a) (4 points) **Inference; Prediction**, respectively.

(b) (4 points)

$$\mathbb{E}[3X - 2Y] = 3\mathbb{E}[X] - 2\mathbb{E}[Y] = 3(2) - 2(-1) = \boxed{8}.$$

(c) (4 points) **False**: independence implies zero covariance, but the converse can fail; for example, if X is uniform on $\{-1, 0, 1\}$, then $Y = X^2$ is dependent on X while $\text{Cov}(X, Y) = 0$.

(d) (4 points) **B**: $\sum_i (y_i - \hat{y}_i)^2$.

(e) (4 points) **Confidence interval**: We are 95% confident that the population mean exam score for students studying 5 hours per week lies between 75 and 83 points.

Prediction interval: Under the regression model, $[60, 98]$ is a 95% prediction interval for the exam score of one new student studying 5 hours per week.

(f) (4 points) $\boxed{3}$ indicators are needed: one for each nonbaseline mode, with the baseline represented by all three indicators being zero and its mean represented by the intercept.

(g) (4 points) **No**: given the email's features, the model estimates an 80% probability that it is spam, which does not guarantee its actual label.

(h) (4 points) **Generative**.

(i) (4 points) **False positive; False negative**, respectively.

(j) (4 points) The factor is $\boxed{2}$, since

$$\frac{p(x+1)/(1-p(x+1))}{p(x)/(1-p(x))} = e^{\log 2} = 2.$$

Problem 2 (20 points). Learning a coin's type**(a) (10 points)** By total probability and Bayes' theorem,

$$P(A) = \frac{1}{2} \cdot \frac{3}{4} + \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{2}, \quad P(T=H | A) = \frac{P(A | T=H)P(T=H)}{P(A)} = \frac{(3/4)(1/2)}{1/2} = \boxed{\frac{3}{4}}.$$

The $\boxed{H\text{-type}}$ coin is more likely after the first head.**(b) (10 points)** Conditional independence given T gives

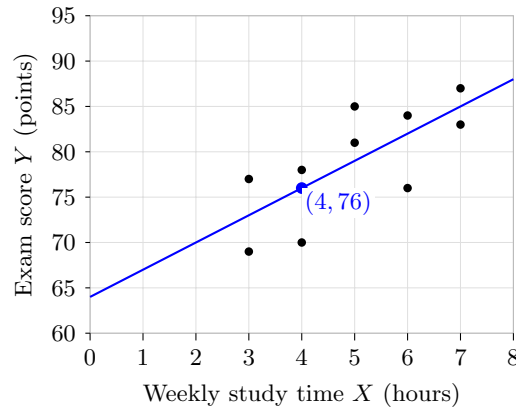
$$P(B) = \frac{1}{2} \cdot \frac{3}{4} + \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{2}, \quad P(A \cap B) = \frac{1}{2} \left(\frac{3}{4}\right)^2 + \frac{1}{2} \left(\frac{1}{4}\right)^2 = \frac{5}{16}.$$

Therefore,

$$P(B | A) = \frac{P(A \cap B)}{P(A)} = \frac{5/16}{1/2} = \boxed{\frac{5}{8}}.$$

Observing the first head increases the probability of the second from $1/2$ to $5/8$, so A and B are not independent without conditioning on T .**Problem 3 (30 points + 5 bonus). Study time and exam scores****(a) (10 points)**

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{60}{20} = 3, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 79 - 3(5) = 64.$$

Thus $\boxed{\hat{y} = 64 + 3x}$, and $\boxed{\hat{y}_{\text{new}} = 64 + 3(4) = 76}$.**(b) (10 points)** Using the separate 60-student sample,

$$\hat{\beta}_1 \pm 2.00 \text{ SE}(\hat{\beta}_1) = 2.5 \pm 2.00(0.5) = \boxed{[1.5, 3.5]}.$$

We are 95% confident that each additional hour of weekly study is associated with an increase of 1.5 to 3.5 points in the population mean exam score.

(c) (10 points) The raw difference and the fitted-model decomposition agree:

$$74 - 82 = -8 = 4(3 - 8) + 12 = -20 + 12.$$

No: participants study 5 fewer hours on average, accounting for a 20-point lower fitted score, while the model predicts a 12-point higher score for participants at the same study time. Because participation was not randomly assigned, these associations do not establish a causal effect of the program.

(d*) (5 bonus points) The augmented predictor gives

$$a_0 + a_1x + a_2z(x) = a_0 + a_1x + a_2(b_0 + b_1x) = (a_0 + a_2b_0) + (a_1 + a_2b_1)x.$$

This is exactly the original class of affine functions, and the presence of two distinct training inputs makes its least-squares minimizing function unique, so every refit predicts $g(x)$ for every training or new x , although the three coefficients need not be unique. Consequently, the minimum training RSS is unchanged and the additional predictor causes no additional overfitting.

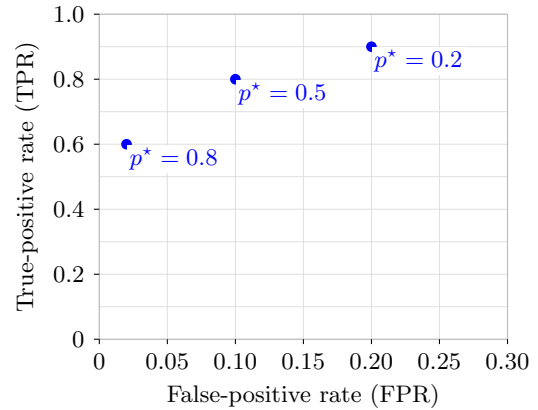
Problem 4 (30 points + 5 bonus). Reading a spam-filter study

(a) (10 points) **Logistic regression** directly models $P(Y = 1 | X)$ through a logistic link, giving probabilities in $(0, 1)$ without specifying feature distributions within each class.

(b) (10 points) There are 100 actual spam emails and 900 actual nonspam emails, so

$$\text{TPR} = \frac{TP}{100}, \quad \text{FPR} = \frac{FP}{900}.$$

p^*	FPR	TPR
0.2	$180/900 = 0.20$	$90/100 = 0.90$
0.5	$90/900 = 0.10$	$80/100 = 0.80$
0.8	$18/900 = 0.02$	$60/100 = 0.60$



Choose $p^* = 0.5$: thresholds 0.5 and 0.8 meet $\text{FPR} \leq 0.10$, and 0.5 has the higher TPR, 0.80.

(c) (10 points) Let F be the event that an email is flagged. In the new stream,

$$\begin{aligned} P(Y = 1 | F) &= \frac{P(F | Y = 1)P(Y = 1)}{P(F | Y = 1)P(Y = 1) + P(F | Y = 0)P(Y = 0)} \\ &= \frac{0.80(0.01)}{0.80(0.01) + 0.10(0.99)} = \boxed{\frac{8}{107} \approx 0.075}. \end{aligned}$$

The claim is false: the smaller spam prevalence lowers this probability, from $80/170 = 8/17 \approx 0.471$ in the evaluation set to about 0.075 in the new stream.

(d*) (5 bonus points) Conditional on $X = x$, the errors are

$$R_A(x) = P(\hat{Y}_A \neq Y | X = x) = 1 - p,$$

$$R_B(x) = P(\hat{Y}_B \neq Y | X = x) = p(1 - p) + (1 - p)p = 2p(1 - p),$$

where the products for Rule B use the stated conditional independence of its draw and the true label. Since

$$R_B(x) - R_A(x) = (1 - p)(2p - 1) > 0 \quad \text{for } \frac{1}{2} < p < 1,$$

Rule A has strictly lower conditional probability of error.