Math 532 • Regression 2 Name _____

Midterm Solutions, March 15, 2023 Instructor: Dikran Karagueuzian

1. Suppose that we are doing a retrospective study of a disease. Let D be the event that an individual has the disease, $\bar{D}$ be the event that an individual does not have the disease, and I the event that an individual is included in the study. Let $\pi_0 = P(I \mid D)$, $\pi_1 = P(I \mid \bar{D})$.

   In our study we wish to account for the effect of certain variables. For an individual patient we denote these variables collectively by x. Let $p^*(x)$ be the conditional probability that a patient with these variable values has the disease given that he or she was included in the study, and let $p(x)$ be the unconditional probability that he or she has the disease.

   Bayes' Formula gives:

$$p^*(x) = \frac{\pi_1 p(x)}{\pi_1 p(x) + \pi_0(1 - p(x))}$$

   (a) (5 points) What is the definition of the logit or log-odds function $\text{logit}(p)$?

   Solution:

$$\text{logit}(p) = \log(\frac{p}{1-p})$$

   (b) (15 points) Find $\text{logit}(p(x))$ in terms of $\text{logit}(p^*(x))$, using the Bayes' Formula equation above.

   Solution:

$$p^*(x) = \frac{\pi_1 p(x)}{\pi_1 p(x) + \pi_0(1 - p(x))} \qquad \text{Bayes' Formula}$$

$$\text{logit}(p^*(x)) = \text{logit}\big(\frac{\pi_1 p(x)}{\pi_1 p(x) + \pi_0(1 - p(x))}\big)$$

$$\text{logit}(p^*(x)) = \log\left[\frac{\frac{\pi_1 p(x)}{\pi_1 p(x)+\pi_0(1-p(x))}}{\frac{\pi_0(1-p(x))}{\pi_1 p(x)+\pi_0(1-p(x))}}\right] \qquad \text{definition of logit}$$

$$\text{logit}(p^*(x)) = \log\big(\frac{\pi_1 p(x)}{\pi_0(1 - p(x))}\big)$$

$$\text{logit}(p^*(x)) = \log(\frac{\pi_1}{\pi_0}) + \text{logit}(p(x)) \qquad \text{property of log}$$

$$\text{logit}(p(x)) = \text{logit}(p^*(x)) - \log(\frac{\pi_1}{\pi_0})$$

(c) (10 points) What does this tell us about the effect of the variables x in retrospective studies? (Assume that we do not know π_1 or π_0.)

Solution: We can see the relative effect of the variables, but not the absolute effect. That is, we might know that patients who have systolic blood pressure 10 points higher tend to have log-odds of having the disease larger by 2. But we would not know the absolute log-odds of having the disease for the patient population as a whole or any subgroup with particular values of our predictor variables.

2. Assume that you have a linear regression model for the price of Honda Accords (a model of car) in terms of age in years, and miles on the odometer (measured in units of $10,000$). (So, a brand new car is 0 years old and has 0 miles, while a 3-year old car with $40,000$ miles has values of 3 and 4.) This model is given by the equation:

$$\text{price} = \alpha_0 + \alpha_1 \cdot \text{miles} + \alpha_2 \cdot \text{age} + \epsilon,$$

where ϵ is a normally distributed error. You fit this model and obtain coefficients $\widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\alpha}_2$.

Cars that are older and/or have more miles tend to have lower prices.

Cars that are older tend to have more miles on the odometer, and vice versa.

(a) (6 points) What sign(s) do you expect $\widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\alpha}_2$ to have?

Solution: We expect that $\widehat{\alpha}_0$ is positive: new cars cost something. We expect that $\widehat{\alpha}_1, \widehat{\alpha}_2$ are negative: cars that are older and/or have more miles tend to have lower prices.

You now fit a new model with age omitted, i.e.:

$$\text{price} = \beta_0 + \beta_1 \cdot \text{miles} + \epsilon,$$

where ϵ is a normally distributed error, and obtain coefficients $\widehat{\beta}_0, \widehat{\beta}_1$.

(b) (8 points) You now fit a new model with age omitted, i.e.:

$$\text{price} = \beta_0 + \beta_1 \cdot \text{miles} + \epsilon,$$

where ϵ is a normally distributed error, and obtain coefficients $\widehat{\beta}_0, \widehat{\beta}_1$.
Do you expect that $\widehat{\beta}_1 \approx \widehat{\alpha}_1$, $\widehat{\beta}_1 < \widehat{\alpha}_1$, or $\widehat{\beta}_1 > \widehat{\alpha}_1$?

Solution: We expect that $\widehat{\beta}_1 < \widehat{\alpha}_1$.

(c) (8 points) Give a reason for your answer in the previous part.

Solution: This is because $\widehat{\alpha}_1$ is negative, $\widehat{\alpha}_2$ is negative, and the correlation of age and miles is positive, so miles acts as a proxy for age. Note also that the order of magnitidue of age and miles is the same in the given units, so the effect is significant.

3. You fit a logistic regression model in R; the summary output is given below:

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -360.61  195972.86 -0.0018   0.9985
Sepal.Width    -110.13   55361.50 -0.0020   0.9984
Sepal.Length    131.79   64576.99  0.0020   0.9984

n = 100 p = 3
Deviance = 0.00000 Null Deviance = 138.62944 (Difference = 138.62944)
```

Some facts about this output indicate a phenomenon we have studied.

(a) (5 points) What is the name of this phenomenon?

Solution: "Perfect Separation", "Linear Separability", and "Existence of a separating hyperplane" are all reasonable answers.

(b) (10 points) What, specifically, in this output indicates that this phenomenon is occurring?

Solution: The standard errors are all very large. The coefficient estimates are all large. Nothing is significant.

(c) (5 points) What are some circumstances in which this phenomenon frequently occurs?

Solution: It often happens that there is a separating hyperplane when the number of data points is of the same order of magnitude as the number of predictor variables. (That's $n < 10p$ in the usual notation.)

(d) (5 points) What could you do in such circumstances to improve your model?

Solution: You could use a regularized regression package, like `brglm`. You could get more data. You could use another binary classification method.

4. (15 points) Suppose we have data on a response Y and predictor variables $X_1, \ldots, X_p$. Consider a model that transforms the response Y to make a linear regression model

$$\log(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

What are some of the differences between this and a Poisson GLM for Y?

Solution: The Poisson GLM assumes that Y is Poisson and that $\log(E[Y]) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$. The model above assumes that Y is real-valued and that $E[\log(Y)] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$. Since Poisson random variables are discrete and not continuous, and $E[\log(Y)] \neq \log(E[Y])$, these are differences.

Also, the model above assumes that the variance of the response $\log(Y)$ is the same whatever the values of $X_1, \ldots, X_p$. In the Poisson GLM, the variance of Y is the same as the mean, which changes with the values of $X_1, \ldots, X_p$.

Other answers are also possible.

5. You fit a logistic regression with a command in R.

   (a) (5 points) What command is it that you use and what options to this command do you use?

   Solution: Use the `glm` command with the `family=binomial` option.

   (b) (5 points) How does this change if you wish to fit a probit regression instead?

   Solution: Use the `glm` command with the `family=binomial` option and the `link=probit` option.

   (c) (5 points) What about a Poisson regression instead of a logistic regression?

   Solution: Change `family=binomial` to `family=poisson`.

6. Suppose that

$$Y = \alpha + \beta X + \epsilon$$

where ϵ is a normally distributed error, i.e. this model is exactly true.

Assume that you do not have access to the true values of X but only to $X' = X + u$, where u is a normally distributed error with mean 0 and variance σ^2 which is independent of X, Y, ϵ.

You now fit a regression

$$Y = \alpha + \beta X' + \epsilon.$$

Recall that your estimate $\widehat{\beta} = \frac{\mathrm{Cov}(Y,X')}{\mathrm{Var}(X')}$. (You learned this fact about linear regression last semester.)

(a) (15 points) Compute $\frac{\mathrm{Cov}(Y,X')}{\mathrm{Var}(X')}$ in terms of $\frac{\mathrm{Cov}(Y,X)}{\mathrm{Var}(X)}$.

Solution:

$$\begin{aligned}
\frac{\mathrm{Cov}(Y,X')}{\mathrm{Var}(X')} &= \frac{\mathrm{Cov}(Y,X+u)}{\mathrm{Var}(X+u)} && \text{definition of } X' \\
&= \frac{\mathrm{Cov}(Y,X)+\mathrm{Cov}(Y,u)}{\mathrm{Var}(X)+\mathrm{Var}(u)} && \text{indep of } u \text{ and bilinearity of Cov} \\
&= \frac{\mathrm{Cov}(Y,X)+0}{\mathrm{Var}(X)+\sigma^2} && \text{indep of } u \text{ and definition of } \sigma^2 \\
&= \frac{\mathrm{Cov}(Y,X)}{\mathrm{Var}(X)} \cdot \frac{\mathrm{Var}(X)}{\mathrm{Var}(X)+\sigma^2}
\end{aligned}$$

(b) (10 points) How does the expected value of your estimate $\widehat{\beta}$ compare to the true value β?

Solution: The computation above shows that

$$\widehat{\beta} = \beta \cdot \frac{\mathrm{Var}(X)}{\mathrm{Var}(X)+\sigma^2},$$

so that $\widehat{\beta}$ is β times a multiplicative factor which is less than one, since variance is positive. In other words, $\widehat{\beta}$ has smaller absolute value than β, i.e. $\widehat{\beta}$ is biased towards zero. This is called "Attenuation Bias".