STA 2101/442 Assignment 8[1]

1. In this question you will develop a prediction interval (not a confidence interval) for Y_{n+1}.

   (a) What is the distribution of $Y_{n+1} - \widehat{Y}_{n+1} = Y_{n+1} - \mathbf{x}_{n+1}^\top \widehat{\boldsymbol{\beta}}$? Show your work. Your answer includes both the expected value and the variance.

   (b) Now standardize the difference to obtain a standard normal.

   (c) Divide by the square root of a chi-squared random variable, divided by its degrees of freedom, and simplify. Call it T. Compare your answer to a slide from lecture. How do you know that numerator and denominator are independent?

   (d) Using your result, derive the $(1-\alpha) \times 100\%$ prediction interval for Y_{n+1}.

2. Recall from lecture the logistic regression in which the response variable is passing a first-year calculus course or not. Here is output for a model with just one explanatory variable, High School Grade Point Average.

```
> mod = glm(passed ~ hsgpa, family=binomial); summary(mod)

Call:
glm(formula = passed ~ hsgpa, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5152  -1.0209   0.4435   0.9321   2.1302

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -15.21013    1.99832  -7.611 2.71e-14 ***
hsgpa         0.19773    0.02548   7.759 8.56e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 530.66  on 393  degrees of freedom
Residual deviance: 443.43  on 392  degrees of freedom
AIC: 447.43

Number of Fisher Scoring iterations: 4
```

```
> V = vcov(mod); V # Approximate asymptotic covariance matrix of beta-hat
             (Intercept)         hsgpa
(Intercept)  3.99328951 -0.0508420333
hsgpa       -0.05084203  0.0006494617
```

For a student with a High School grade point average of 70%,

(a) Give a point estimate of the probability that the student will pass the course. The answer is a number. Use your calculator, as on the final exam.

(b) Using the delta method, obtain a 95% confidence interval for the probability that a new (unobserved) student with this GPA will pass the course. The answer is a pair of numbers, calculated from the output above. My lower confidence limit is 0.127.

3. To predict a binary response based on a logistic regression model, it makes sense to produce a $\widehat{\pi}$ (estimated probability of $y = 1$). One would predict Yes if $\widehat{\pi} > 0.5$, and No if $\widehat{\pi} < 0.5$. A prediction interval would be nice too, but it seems out of reach (to me, at least) because the response is either zero or one.

   On the other hand, we can be more confident in some predictions than others. In Question 2, we can be pretty confident that a student with a HS GPA of 70 will not pass, since the confidence interval did not include 0.5.

   This suggests testing the null hypothesis that $\pi = \frac{1}{2}$ for a particular set of x values, and only make a *confident* prediction if H_0 is rejected. (I'm just making this up, but it seems reasonable.)

   (a) One could do it with the confidence interval for π, but there is an easier way that does not require the delta method. Why is $H_0 : \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} = 0$ a good null hypothesis?

   (b) Using the output from Question 2, Obtain a Z statistic for testing $H_0 : \boldsymbol{\beta}^\top \mathbf{x} = 0$. The answer is a number. What do you conclude?

4. Employees at Canadian corporations filled out questionnaires about their jobs. Questionnaires employed 5-point scales, where 5 indicates the highest level of the trait or opinion being assessed (like job satisfaction) and 1 indicating the lowest level. The wording of the questions was varied so that sometimes a 1 indicated higher satisfaction (for example, strong disagreement with "I hate my job."), but the numbers were switched around so that in the data file, larger numbers always indicate more. Data consist of answers to

   - Ten questions about committment to the organization, with higher numbers indicating more committment.
   - Five questions about relations with colleagues at work, with higher numbers indicating better relations.
   - Twelve questions about relations with magnagement, in particular the respondent's immediate boss. Higher numbers indicate better relations.
   - Six questions about fair opportunities for advancement, with higher numbers indicating more fairness.
   - Four questions about job satisfaction, with higher numbers indicating more satisfaction.
   - Three questions about senior management's committment to diversity, with higher numbers indicating more committment. These seem to be on a six-point scale instead of five.
   - Gender: 0=Male, 1=Female
   - Visible Minority status: 0=No, 1=Yes
   - Education level, numbered 1-7. The exact meanings of the numbers are unknown, but surely higher numbers must indicate more education, mostly.
   - Marital status: 1=never married, 2=married, 3=divorced or separated, 4=widowed.
   - Age in years
   - Born outside Canada: 0=No, 1=Yes

   There are two data sets of size n=500, randomly sampled from around 16,000 questionnaires. These are real data. The data sets are in Excel spreadsheets:

   - http://www.utstat.toronto.edu/ brunner/data/legal/DiversityExplore.xlsx
   - http://www.utstat.toronto.edu/ brunner/data/legal/DiversityReplic.xlsx

   The goal is to develop a predictive model for expressed commitment to the organization; that's the sum of 10 items. Let's stick to least-squares regression this first time. Rather than doing it the way I did in lecture, please impute any missing explanatory variable values. Develop your model on the exploratory sample, and then try it out on the replication sample. We will say that a predictive model is better when the mean absolute error of prediction is small. I recommend the `xlsx` and `DMwR` packages for this.