STA 114: Statistics

Lab 10

Revisiting HW 10 #3

In HW 10, #3 we looked at data $X = (X_1, X_2)$ modeled as $X \sim \mathsf{Multinomial}(n, p)$, $p \in \Delta_2$ and wanted to test $H_0 : p = p_0$ for some fixed $p_0 = (p_{01}, p_{02}) \in \Delta_2$. We talked about two different ways of doing this. First is the usual Pearson's chi-square test and the second is the ML test for the equivalent problem: $X_1 \sim \mathsf{Binomial}(n, p_1)$, $p_1 \in (0, 1)$ and test $H_0 : p_1 = p_{01}$ against $H_1 : p_1 \neq p_{01}$. We saw that the two tests are similar and can be described as:

Size-α Pearson's test. Reject H_0 if $Q(x) = \frac{(x_1 - np_{01})^2}{np_{01}(1-p_{01})} > F_1^{-1}(1-\alpha) = z(\alpha)^2$

Size-α binomial ML test. Reject H_0 if $T(x) = \frac{(x_1 - np_{01})^2}{n(x_1/n)(1-x_1/n)} > z(\alpha)^2$.

We'll assess how similar these two tests really are. Note that the p-value based on Pearson's tests is $1 - F_1(Q(x))$ whereas the same based on the binomial ML tests is $1 - F_1(T(x))$.

Task 1. Fix $n = 500$ and $p_0 = (0.4, 0.6)$. Generate an x from $\mathsf{Multinomial}(n, p_0)$ [it suffices to get $x_1 \sim \mathsf{Binomial}(n, p_{01})$ and set $x_2 = n - x_1$]. Calculate the two p-values for this data under the two kinds of tests. Now repeat this 100 times. Make a plot of these 100 pairs of p-values and compare them with the 45 degree line. Are the two types of tests behaving similarly for this simulation setting?

Task 2. Continue using $n = 500$ and $p_0 = (0.4, 0.6)$ but now generate an x from $\mathsf{Multinomial}(n, (0.5, 0.5))$ and get the two p-values for this data. Repeat this 100 times and make a plot of the 100 pair of p-values. Compare against the 45 degree line. Are the two tests still behaving similarly?

Task 3. Repeat the above task but generate your x's from $\mathsf{Multinomial}(n, (0.9, 0.1))$.

Task 4. Which type of tests appears to have smaller type II error probabilities? Can you make a precise statement?

Performing simple linear regression on R

Next we'd look at how to perform simple linear regression on R. To start with load the dataset "highway" from R package "alr3". This dataset contains observations from 39 segments of a highway

about on accident rate (`Rate`) and physical characteristics of the road segment including speed limit (`Slim`), number of access points (`Acpt`), etc. We'll work with accident rate as the response and speed limit as the explanatory variable.

```
library(alr3)
data(highway)
y <- highway$Rate
x <- highway$Slim
```

If you can't load the data from the library, simple copy and paste the following

x	55, 60, 60, 65, 70, 55, 55, 55, 50, 50, 60, 50, 50, 60, 55, 60, 60, 50, 55, 60, 55, 60, 50, 60, 40, 45, 55, 55, 45, 60, 45, 55, 55, 55, 55, 50, 55, 60, 55
y	4.58, 2.86, 3.02, 2.29, 1.61, 6.87, 3.85, 6.12, 3.29, 5.88, 4.2, 4.61, 4.8, 3.85, 2.69, 1.99, 2.01, 4.22, 2.76, 2.55, 1.89, 2.34, 2.83, 1.81, 9.23, 8.6, 8.21, 2.93, 7.48, 2.57, 5.77, 2.9, 2.97, 1.84, 3.78, 2.76, 4.27, 3.05, 4.12

TASK 5. Get n, $\bar{x}$, s_x^2, s_{xy} and $\bar{y}$ from the data. Recall $s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$ and $s_{xy} = \frac{1}{n-1}\sum_{i=1}^n y_i(x_i - \bar{x})$.

TASK 6. Calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ based on these numbers. Also calculate $\hat{\sigma}$. Recall $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$.

TASK 7. Now use the `lm()` function of R to perform the data analysis and check them against your answers. You can use the following codes

```
hwy.lm <- lm(y ~ x)
print(hwy.lm)
summary(hwy.lm)
```

The `summary()` function prints a lot of details in addition to the estimated values of the intercept and the slope. In particular "`Residual standard error`" gives $\hat{\sigma}$.

TASK 8. Extract the residuals $e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$, $i = 1, \cdots, n$ by using

```
resid <- hwy.lm$resid
```

and draw a histogram of these residuals (with `freq = FALSE`). Overlay the pdf of $\mathsf{Normal}(0, \hat{\sigma}^2)$ on the histogram.

I'd leave you one thing to think about. Why is the slope estimated negative? Should increasing the speed limit decrease highway accident rate? Or is there something more subtle going on for this dataset? (Think about which segments of a highway are likely to have lower speed limits.)