STATS 191: Homework Assignment 2

Dr. Pratheepa Jeganathan

10/04/2019

You may discuss homework problems with other students, but you have to prepare the written assignments yourself.

Please combine all your answers, the computer code and the figures into one PDF file, and submit a copy to gradescope.

Grading scheme: $\{0, 1, 2\}$ **points per question, total of 54.**

Due date: 11:59 PM October 11, 2019 (Friday evening).

Question 1

In a recent, exciting, but also controversial Science article, Tomasetti and Vogelstein attempt to explain why cancer incidence varies drastically across tissues (e.g. why one is much more likely to develop lung cancer rather than pelvic bone cancer). The authors show that a higher average lifetime risk for a cancer in a given tissue correlates with the rate of replication of stem cells in that tissue. The main inferential tool for their statistical analysis was a simple linear regression, which we will replicate here.

You can download the dataset as follows:

```
tomasetti = read.csv("https://stats191.stanford.edu/data/Tomasetti.csv")
head(tomasetti)
```

```
##                                           Type   Risk     Lscd
## 1                       Acute myeloid leukemia 0.0041 1.30e+11
## 2                         Basal cell carcinoma 0.3000 3.55e+12
## 3                 Chronic lymphocytic leukemia 0.0052 1.30e+11
## 4                    Colorectal adenocarcinoma 0.0480 1.17e+12
## 5            Colorectal adenocarcinoma with FAP 1.0000 1.17e+12
## 6 Colorectal adenocarcinoma with Lynch syndrome 0.5000 1.17e+12
##         Cluster
## 1   Replicative
## 2 Deterministic
## 3   Replicative
## 4 Deterministic
## 5 Deterministic
## 6 Deterministic
```

The dataset contains information about 31 tumour types. The `Lscd` (Lifetime stem cell divisions) column refers to the total number of stem cell divisions during the average lifetime, while `Risk` refers to the lifetime risk for cancer of that tissue type.

1. Fit a simple linear regression model to the data with `log(Risk)` as the response variable and `log(Lscd)` as the predictor variable.
2. Plot the estimated regression line and the data.
3. Add upper and lower 95% prediction bands for the regression line on the plot, using `predict`. That is, produce one line for the upper limit of each interval over a sequence of densities, and one line for the lower limits of the intervals.

4. Interpret the above bands at a `Lscd` of 10^{10}.
5. Add upper and lower 95% confidence bands for the regression line on the plot, using `predict`. That is, produce one line for the upper limit of each interval over a sequence of densities, and one line for the lower limits of the intervals.
6. Interpret the above bands at a `Lscd` of 10^{10}.
7. Test whether the slope in this regression is equal to 0 at level $\alpha = 0.05$. State the null hypothesis, the alternative, the conclusion and the p-value.
8. What are assumptions you made for question (7).
9. Give a 95% confidence interval for the slope of the regression line.
10. Interpret your interval in (9).
11. Report the R^2 of the model.
12. Report the adjusted R^2 of the model.
13. Report an estimate of the variance of the errors in the model.
14. Provide an interpretation of the R^2 you calculated above.
15. According to a Reuters article "Plain old bad luck plays a major role in determining who gets cancer and who does not, according to researchers who found that two-thirds of cancer incidence of various types can be blamed on random mutations and not heredity or risky habits like smoking." Is this interpretation of R^2 correct?

Question 2

From our textbook **CH** page 51, Exercie 2.9.

Let Y and X denote the labor force participation rate of women in 1972 and 1968, respectively, in each of 19 cities in the United States. The regression output for this data set is shown in the following table. It was also found that SSR = .0358 and SSE = .0544. Suppose that the model $Y = \beta_0 + \beta_1 X + \epsilon$ satifies the ususal regression assumptions.

Variable	Coefficient	s.e	t-Test	p-value
Constant	.203311	.0976	2.08	.0526
X	.656040	.1961	3.35	$< .0038$
–	–	–	–	–
n = 19	$R^2 = .397$	$R_a^2 = .362$	$\hat{\sigma} = .0566$	df = 17

In this table **s.e** is the standard error of the estimate, **t-Test** is the value of the test statistics under the null hypothesis, **p-value** is the p-value of the test.

1. Compute $\text{Var}(Y)$ and $\text{Cov}(Y, X)$.
2. Suppose participation rate of women in 1968 in a given city is 45%. What is the estimated participation rate of women in 1972 for the same city?
3. Suppose further that the mean and variance of the participantion rate of women in 1968 are 0.5 and 0.005, respectively. Construct the 95% confidence interval for the estimate in (2)
4. Construct the 95% confidence interval for the slope of the true regression line β_1.
5. Test the hypothesis: $\text{H}_0 : \beta_1 = 1$ versus $\text{H}_a : \beta_1 > 1$ at the 5% significance level.

6. Compute the R^2 for this simple linear regression.
7. If X and Y were reversed in the above regression, what would you expect R^2 to be?

Question 3

Power is an important quantity in many applications of statistics. This question investigates the power of a test in simple linear regression. In a simple linear regression setting, suppose the true slope of the regression line is β_1 and the true intercept is β_0. If we assume σ **is known**, then we can test

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_a : \beta_1 \neq 0$$

using

$$Z = \frac{\hat{\beta}_1 - 0}{SD(\hat{\beta}_1)}$$

where $SD(\hat{\beta}_1)$ is the standard deviation of our estimator $\hat{\beta}_1$. (We are using Z here instead of T to avoid complication of the degrees of freedom – just imagine our sample size n was really large. In this case our estimate of variability $SE(\hat{\beta}_1)$ is replaced by the true standard deviation $SD(\hat{\beta}_1)$, i.e. we have swapped $\hat{\sigma}^2$ with σ^2.)

For a fixed significance level α, the power of this test is a function of the true value β_1 as well as the accuracy of our estimate $SD(\hat{\beta}_1)$. The power is defined as

$$P_{(\beta_0,\beta_1)}\left(H_0 \text{ is rejected} | H_0 \text{ is false}\right).$$

That is, the probability we reject the null hypothesis as a function of (β_0, β_1). Actually, the power will generally not depend on β_0 in this model, so it is really a function of β_1 (and $SD(\hat{\beta}_1)$).

As we change the true β_1, the probability we reject H_0 changes: if the true value of β_1 is much larger than 0 relative to $SD(\hat{\beta}_1)$ then we are very likely to reject H_0.

1. What rule would you use to determine whether or not you reject H_0 at level $\alpha = 0.05$ (write down the rule in terms of the rejection region).
2. When H_0 is false, what is the distribution of our test statistic $Z = \dfrac{\hat{\beta}_1}{SD(\hat{\beta}_1)}$? Show that the distribution depends only on the value $\dfrac{\beta_1}{SD(\hat{\beta}_1)}$. We call this quantity the non-centrality parameter or signal to noise ratio (SNR).
3. For $\alpha = .05$, plot the power of your test as a function of the SNR. (Write down Power $= P_{(\beta_0,\beta_1)}\left(H_0 \text{ is rejected} | H_0 \text{ is false}\right)$ in terms of the distribution of Z when H_0 is false and the critical value).
4. Using the above plot in (3), find out the power when SNR $= 0$. That is, when $\beta_1 = 0$.
5. Using the above plot in (3), roughly how large does the non-centrality parameter (SNR) have to be in order to achieve power of 85%?