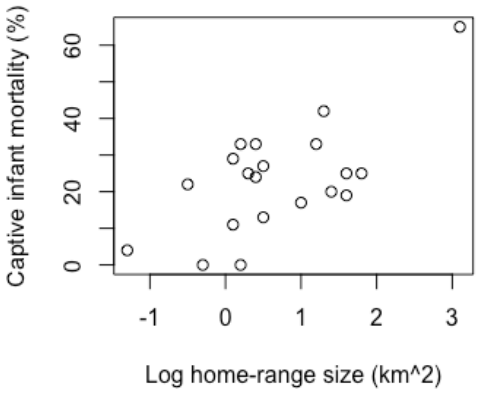
Problem set 3

Due date and time: October 26, 23:30

1. (20 points) Some species seem to thrive in captivity, whereas others are prone to health and behavior difficulties when caged. Maternal care problems in some captive species, for example, lead to high infant mortality. Can these differences be predicted? The following data are measurements of the infant mortality (percentage of births) of 20 carnivore species in captivity along with the log (base-10) of the minimal home-range sizes (in km^2) of the same species in the wild (Clubb and Mason 2003).

Log_{10} home-range size	Captive infant mortality (%)
−1.3	4
−0.5	22
−0.3	0
0.2	0
0.1	11
0.5	13
1.0	17
0.3	25
0.4	24
0.5	27
0.1	29
0.2	33
0.4	33
1.3	42
1.2	33
1.4	20
1.6	19
1.6	25
1.8	25
3.1	65

(a) (3 points) Based on the scatter plot, please describe the relationship between the two variables in words.



(b) (3 points) Based on the output from statistical software R, what are the slope and intercept of the least-squares regression line, with the log of home-range size as the explanatory variable?

```
Call:
lm(formula = mortality ~ log_home_range)

Residuals:
    Min      1Q  Median      3Q     Max
-18.423  -9.688   2.249  10.972  16.810

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      16.370      3.136   5.221 5.77e-05 ***
log_home_range   10.264      2.694   3.810  0.00128 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.38 on 18 degrees of freedom
Multiple R-squared:  0.4465,    Adjusted R-squared:  0.4157
F-statistic: 14.52 on 1 and 18 DF,  p-value: 0.001281
```

(c) (10 points) Does home-range size in the wild predict the mortality of captive carnivores? Carry out a formal test (please use the information from the R output above). Assume that the species data are independent. Please state the null and alternative hypotheses, specify the statistic and the P-value, and draw the conclusion.

(d) (4 points) Is there any outlier that might change the conclusion in (c)? Why or why not?

1.

(a.) As the home-range size (log form) increases, the captive infant mortality tends to increase. But the pattern needs further analysis, and there may have outliers.

⇒ It seems like there is a positive linear relationship between two variables.

(b.) Intercept: 16.370
Slope: 10.264
} $y = 10.264$ (log home-rang size) + 16.370

↳ positive linear relationship

(c). Null Hypothesis: Slope = 0

(No relationship between home-range size and captive infant mortality)

Alternative Hypothesis: slope ≠ 0

(There is a relationship between home-range size and captive infant mortality)

p-value = 0.00128 < 0.05 ⇒ We can reject null hypothesis

The result implies that there is a statiscally significant relationship between home-range size and the captive infant mortality.

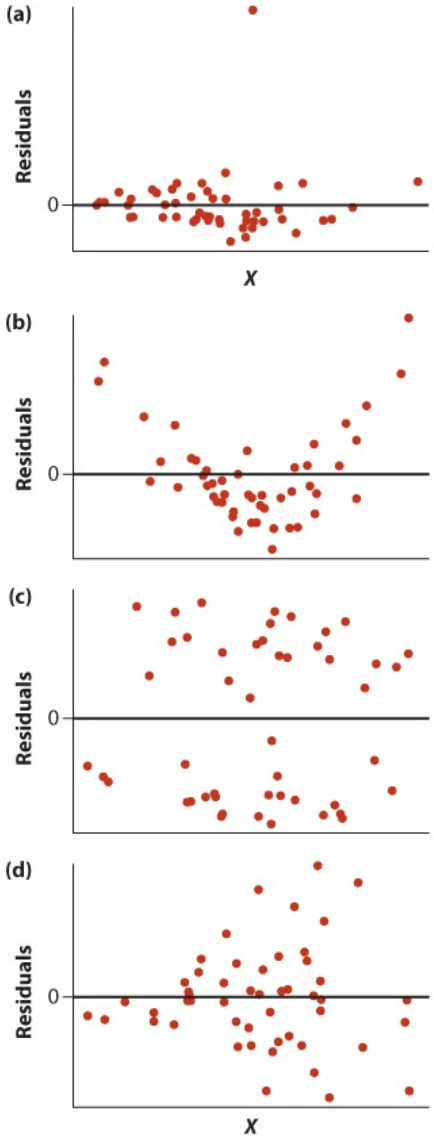
⇒ The home-range size can be a predictor of captive infant mortality.

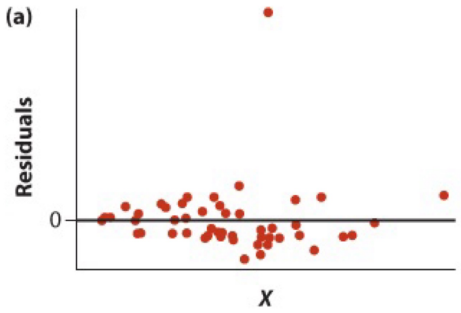
(d). Yes, the data point with a captive infant mortality rate of 65% appears to be an outlier. Besides, 0% mortality rate might also seem like an outlier. Outliers can affect the slope and intercept, making relationship appear stronger or weaker. ⇒ p-value may change

⇒ We could remove the outliers temporarily, and see how p-value would change.

⇒ Plot residual plot to get a better understanding.

2. (20 points) Examine the following residual plots and identify the assumption(s) of linear regression that is (are) violated in each of the following residual plots.





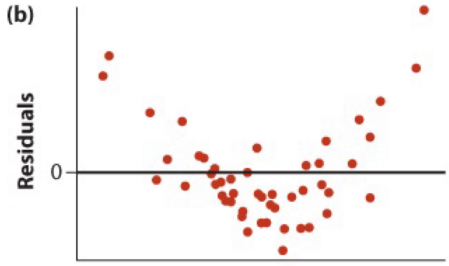
Residuals are consistent as X increases

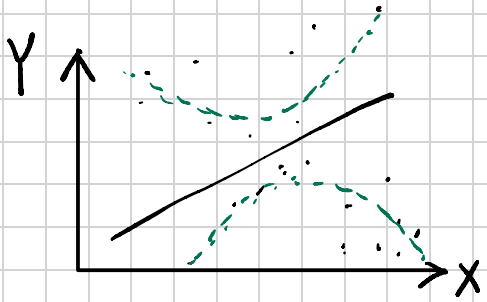
↑

It seems like most of the datapoints are close to the zero line

⇒ Suggesting a good linear fit and equal variance of residuals

However, there is an outlier at middle value of X ⇒ Further analysis is required

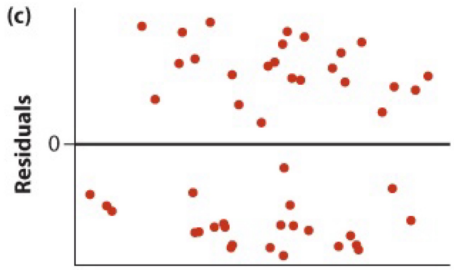




There is a concave curve ⇒ X and Y are non-linear

Residual values seem to maintain a consistent spread but slightly increase at the tail of the curve ⇒ It might violate equal variance assumption

(Further analysis is needed)

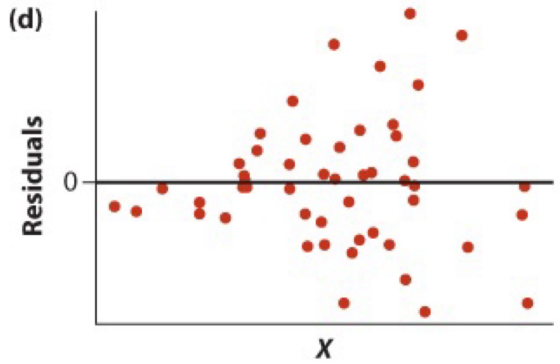


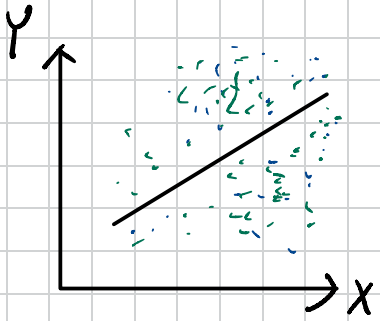
Residuals are scattered but distant from the zero line

⇒ X and Y might have a weak linear relationship

The spread of residuals looks consistent across different X values

⇒ It might follow equal variance assumption

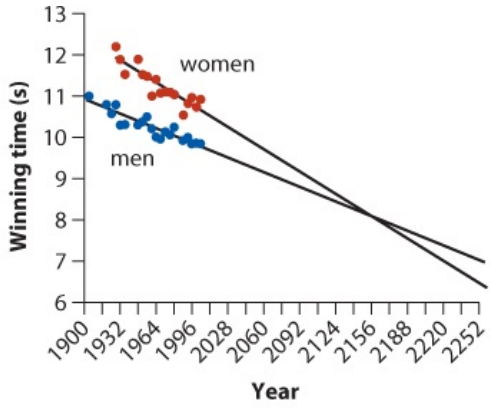




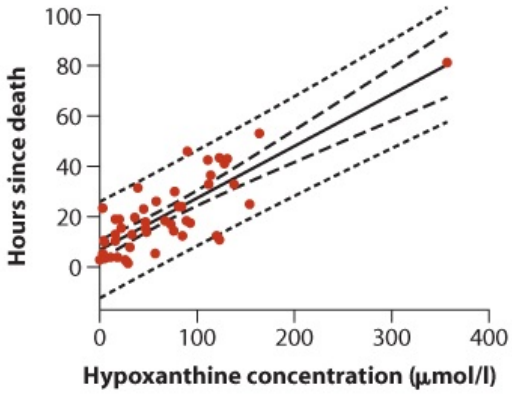
It shows that residuals tend to cluster at the middle value of X and spread out towards the end ⇒ X and Y are non-linear (This regression line might not be effective)

The spread of residuals is not consistent (cluster in the middle & increased spread at tail) ⇒ Violates equal variance assumption

3.(5 points) The slopes of the regression lines on the following graph show that the winning Olympic 100-m sprint times for men and women have been getting shorter and shorter over the years, with a steeper trend in women than in men (the graph is modified from Tatem et al. 2004). If trends continue, women are predicted to have a shorter winning time than men by the year 2156. What cautions should be applied to this conclusion? Explain.



4. (20 points) James et al.(1997) demonstrated that the chemical hypoxanthine in the vitreous humour (the colorless jelly filling the eye) shows a postmortem linear increase in concentration with time since death. This suggests that hypoxanthine concentration might be useful in predicting time of death when it is unknown. The following graph shows measurements collected by the researchers on 48 subjects whose time of death was known. The regression line, the 95% confidence bands, and the 95% prediction interval are included on the graph.



(a) The data set depicted in the graph includes one conspicuous outlier on the far right. If you were advising the forensic scientists who gathered these data, how would you suggest they handle the outlier?
(b) What do the confidence bands measure?
(c) Are the inner dashed lines the confidence bands or the prediction interval?
(d) If the regression depicted in the graph was to be used to predict the time of death in a murder case, which bands would provide the most relevant measure of uncertainty, the confidence bands or the prediction interval? Why?

3. Cautions:

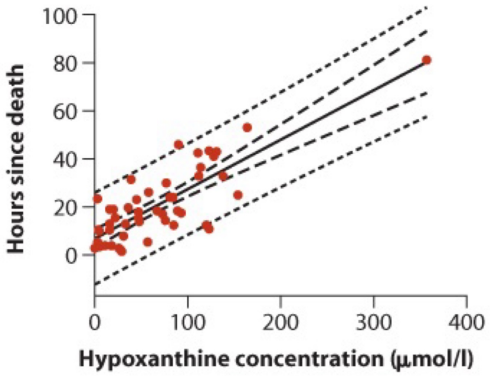
① Biological limits: This prediction only based on the winning time for both genders, it doesn't take biological constraints on human into account. ⇒ The conclusion is not rigorous enough.

② Extrapolation limitation: The extending predictions might far beyond the data range, which could cause more uncertainty. Besides, environment evolves and changes drastically within these years, the prediction should also conclude environmental variables.

more than 100 years

③ External variables: Similar with the previous statements, there are many other variables that could affect winning time, such as training methods, biological features, track, weather...etc. It would be more reliable to fix the requirements or consider more external variables.

4.



(a) The data set depicted in the graph includes one conspicuous outlier on the far right. If you were advising the forensic scientists who gathered these data, how would you suggest they handle the outlier?
(b) What do the confidence bands measure?
(c) Are the inner dashed lines the confidence bands or the prediction interval?
(d) If the regression depicted in the graph was to be used to predict the time of death in a murder case, which bands would provide the most relevant measure of uncertainty, the confidence bands or the prediction interval? Why?

(a.)

Transform the hypoxanthine concentration data into logarithms form and see if the outlier is closer to the rest of distribution. Besides, if the amount of data points are enough, we can just simply remove the outlier.

(b.)

Confidence band: Measure the predicted average time of death at a given value of hypoxanthine concentration.

Prediction interval: Measure the predicted hours since death at a given value of hypoxanthine concentration

(c).

Inner dashed line: Confidence band

Outer dashed line: Prediction interval

(d).

For a specific murder case, prediction interval is most relevant because it provides a range of individual death time at a given concentration of hypoxanthine.

We want to predict death time of a specific person in a murder case.

5. (35 points) Please download the Wine Quality dataset from this link (https://archive.ics.uci.edu/dataset/186/wine+quality) and analyze the data to explore any question of interest. Be sure to clearly state your question of interest and hypothesis and draw conclusions based on the statistical method(s) you choose. You may use any programming language for your analysis, and please include your code when submitting your homework.

Considerations for your analysis:

- When selecting a statistical method, explain why you chose that particular test. Are the assumptions of the chosen method met?
- If not, how might this affect the interpretation of your results?
- Is your analysis focused on prediction or understanding the relationship between variables?

Example questions (not limited to these):

- How can we predict wine quality based on pH value? Should other variables be considered simultaneously in the model?
- Are there significant differences between red and white wines?

When selecting a statistical method, explain why you chose that particular test.
Are the assumptions of the chosen method met?

I chose a classification model because the target variable, wine quality, is ordinal and the relationship between the features and the target is discrete. Since there are 10 features, I think Random Forest would be a suitable option because it can handle high-dimensional data.

The assumptions for Random Forest are met since it is a non-parametric model, it doesn't need to assume normality or equal variance rule. Besides, there are no strong assumptions for features distribution.

Is your analysis focused on prediction or understanding the relationship between variables?

I think it will be mainly focusing on prediction rather than the understanding the relationship between variables. Random Forest is considered a "black box" method, it may be difficult to interpret the underlying relationship between variables.

- How can we predict wine quality based on pH value? Should other variables be considered simultaneously in the model?

Null hypothesis: The pH value alone is sufficient to predict wine quality accurately
Alternative hypothesis: Other variables are significantly contribute to the prediction of wine quality

I have done three analysis:

1. **Predictions by taking all features into account:**
The accuracy score is around 70%, showing that the model performs reasonably well when considering all of the features. The classification report also shows that the model perform better on specific ratings (5, 6, 7) but poor on some classes (3 and 9). This indicates that the classes are imbalance, ratings are mostly around 5-7.

2. **Feature importances:**
Alcohol and volatile acidity are the top two features, with importance scores of 0.132 and 0.103 respectively. This shows that these features have a stronger impact on wine quality than pH value alone. And other features are mostly around 0.7 to 0.9, showing that all of the features are almost equally important.

3. **Predictions by considering pH value alone:**
When using only pH to predict wine quality, the accuracy drops to 43%, implying that pH value an alone is a poor predictor of wine quality. Besides, the classification report also shows low precision, recall and F score for most of the classes.

Final conclusion:

Based on these three analysis results, it is quite obvious that pH alone is not sufficient to predict wine quality accurately. The drop of accuracy from 70% to 40% indicates that other features also play an important role in determining wine quality. Thus, other variables should definitely be considered simultaneously in the model.

(I'm not sure if the null hypothesis can be rejected based on these results)

https://github.com/Annie2305/2024Statistics_and_ML.NTHU_Fall.git

Problem set 3-5