STAT 515 fa 2023 Final Exam

Karl Gregory

- Do not open this exam until told to do so.
- You may have three handwritten sheets of notes out during the exam.
- You have 2.5 hours to work on this exam.
- You may NOT use any kind of calculator.
- If you are unsure what a question is asking for, please ask me for clarification.
- *Good luck, and may the odds be ever in your favor!*

Some survey data collected on the first day of class is appended to this exam.

1. Use the class survey data to investigate whether there is an association between being a CIS major and being a coffee drinker.

   (a) Use the class survey data appended to this exam to fill out the table of observed counts. Include the totals. One major is a missing value (NA); treat this as a non-CIS major.

	Coffee		
Major	Yes	No	Total
CIS			
not CIS			
Total			

The counts are

	Coffee		
Major	Yes	No	Total
CIS	9	7	16
not CIS	11	11	22
Total	20	18	38

   (b) Write down the null and alternate hypotheses which are of interest.

We are interested in testing

H_0: There is no association between being a CIS major and being a coffee drinker.
H_1: There *is* an association between being a CIS major being a coffee drinker.

   (c) The R output for Pearson's Chi-squared test of association is the following:

```
        Pearson's Chi-squared test

data:  data
X-squared = 0.14514, df = 1, p-value = 0.7032
```

From the R output, state your conclusion about the null hypothesis from part (b).

Since the p-value is quite large, there is no reason to reject H_0. For example, if we use significance level $\alpha = 0.05$, we will fail to reject H_0.

(d) Suppose four of the CIS majors who drink coffee had answered that they do not drink coffee. What affect would this have on the p-value in the output of part (c)?

If the data were changed in this way, the data would become The counts are

	Coffee		
Major	Yes	No	Total
CIS	5	11	16
not CIS	11	11	22
Total	16	22	38

These data exhibit a larger difference between the proportion of coffee drinkers in the CIS major versus other majors. This would produce a smaller p-value, because the data would now cast more doubt on the null hypothesis.

2. Students in this class reported on a survey the weight of their keys in grams and their student year —freshman, sophomore, junior, senior, or graduate student. Consider checking whether there is any difference in the mean weight of keys between these groups of students. The R code

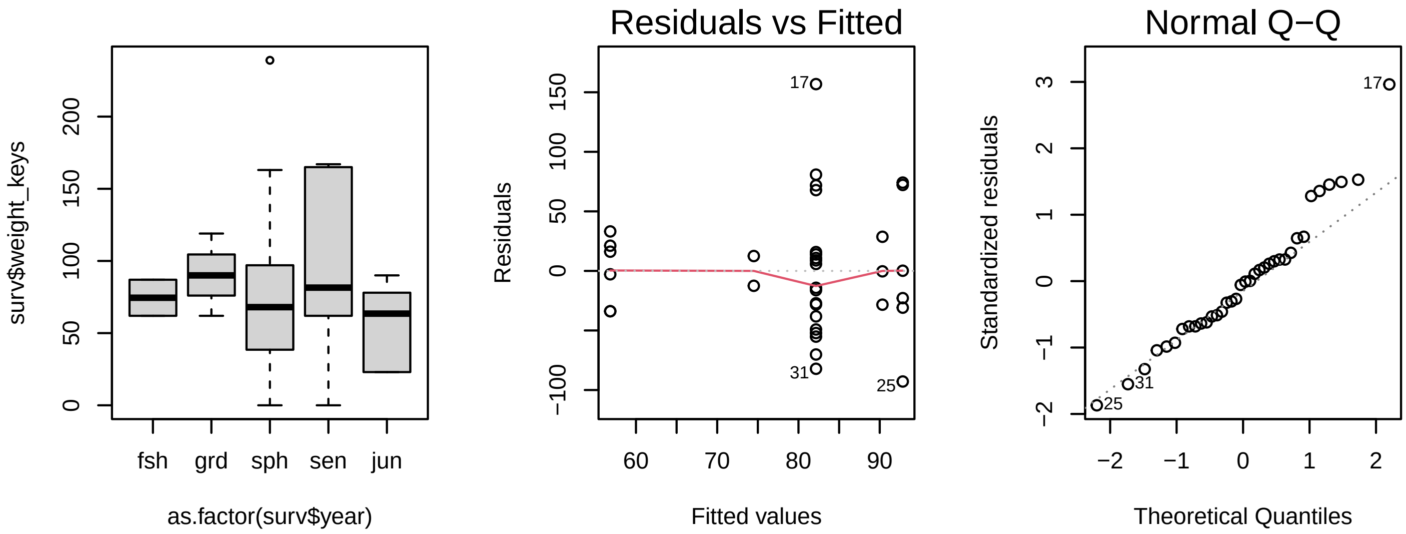
```
lm_out <- lm(surv$weight_keys ~ as.factor(surv$year))
anova(lm_out)
par(mfrow = c(1,3))
boxplot(surv$weight_keys ~ as.factor(surv$year),
        names = c("fsh","grd","sph","sen","jun"))
plot(lm_out,which=c(1,2))
```

prints to the console the output

```
        Analysis of Variance Table

Response: surv$weight_keys
                     Df Sum Sq Mean Sq F value Pr(>F)
as.factor(surv$year)  4   4678  1169.4  0.3953 0.8104
Residuals            31  91713  2958.5
```

and generates the plots



(a) Write down the hypotheses of interest in terms of the group means μ_{fsh}, μ_{grd}, μ_{sph}, μ_{sen}, and μ_{jun}.

The null and alternate hypotheses of interest are

H_0: $\mu_{\text{fsh}} = \mu_{\text{grd}} = \mu_{\text{sph}} = \mu_{\text{sen}} = \mu_{\text{jun}}$
H_1: Not all the means are equal.

(b) Do these data come from a comparative experiment or an observational study? Explain your answer.

The data come from an observational study; in a comparative experiment the subjects or experimental units are randomly assigned to treatment groups. In this study, the experimental units—the students—were not randomly assigned to the groups freshmen, sophomore, junior, senior, and graduate student; rather, they were already in those groups when the data were collected.

(c) The word "Residuals" appears in the output. Explain what the residuals are.

The residuals are the individual response values minus the treatment means. They are the random deviations of the individual response values around the mean of the group to which they belong.

(d) What assumption is one checking when one looks at the residuals versus fitted values plot?

One is checking whether the variances are equal in all the groups.

(e) What assumption is one checking when one looks at the Normal QQ plot of the residuals?

One is checking whether the response values are Normally distributed around the group means.

(f) Give your own diagnosis about whether the assumptions of the analysis are satisfied.

The Normal QQ plot looks fine; the points follow more or less a straight line. The residuals vs fitted values plot, however, exhibits unequal spreads of the residuals across the treatment groups. It seems that the assumption of equal variances might not be satisfied.

(g) The F test statistic for testing whether there is a difference between the means of the groups is a ratio of quantities describing between-group variation and within-group variation. Give the numbers (i) and (ii) from the R output with which the F test statistic is computed as $F = (i)/(ii)$.

This is the mean square of the treatment over the mean square of the error, which is 1169.4/2958.5.

(h) Supposing (this may or may not be the case) that the assumptions of the analysis are satisfied, give your conclusion about your hypotheses in part (a).

Since the p-value is 0.8104, there very little evidence in the data that the students in different years of school have different mean key weights. We fail to reject the null hypothesis.

3. Use the data in the appended class survey to answer this question. Consider the experiment of randomly sampling a single student from the class and letting X be the number of siblings the student has. The numbers of siblings reported by the students in the class are given in the `sibs` column of the survey data set.

   (a) Tabulate the probability distribution of X.

   We obtain the table

x	0	1	2	3	4
$P(X = x)$	2/38	20/38	14/38	1/38	1/38

   (b) Compute the expected value of X. Express your answer as a fraction.

   The expected value is 55/38.

4. Suppose 2% of children in a population have a certain food allergy and suppose there is a test which, if a child does not have the allergy, will give a positive result with probability 0.20; on the other hand, if a child has the allergy, the test will give a positive result with probability 0.90.

   (a) If a child drawn from the population tests positive, give the probability that the child does not have the food allergy. You do not have to simplify your answer.

   It is given that $P(A) = 0.02$ as well as $P(+|A) = 0.90$ and $P(+|A^c) = 0.20$. We have

$$P(A^c|+) = \frac{P(+|A^c)P(A^c)}{P(+|A^c)P(A^c) + P(+|A)P(A)} = \frac{0.20(0.98)}{0.20(0.98) + 0.90(0.02)}$$

   (b) If a child drawn from the population tests negative, give the probability that the child does not have the food allergy. Again, you do not need to simplify your answer.

   We have

$$P(A^c|-) = \frac{P(-|A^c)P(A^c)}{P(-|A^c)P(A^c) + P(-|A)P(A)} = \frac{0.80(0.98)}{0.80(0.98) + 0.10(0.02)}$$

   (c) Are the events that a child tests negative and a that a child has the allergy mutually exclusive?

   No, because $P(- \cap A) = P(-|A)P(A) = 0.10(0.02) \neq 0$.

5. For a random sample $X_1, \ldots, X_n$, recall the formula for the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and for the sample variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$.

   (a) What does S_n^2 describe about a random sample?

   The sample variance S_n^2 describes how spread out the values are.

   (b) Compute S_n^2 for a random sample with the three values: $3, 7, 11$.

   We obtain $\bar{X}_n = 7$ and the $S_n^2 = ((3-7)^2 + (7-7)^2 + (11-7)^2)/2 = 16$.

   (c) Consider S_n^2 for a random sample with the values $6, 7, 8$. Will it be smaller or larger than the value of S_n^2 from part (b)?

   It will be smaller, since the values are less spread out. In fact it will be $S_n^2 = ((6-7)^2 + (7-7)^2 + (8-7)^2)/2 = 1$.

6. Answer each question as TRUE or FALSE.

   (a) The larger the p-value the more implausible the null hypothesis is in light of the data.

   This is **false**. The smaller the p-value the more implausible the null hypothesis is in light of the data.

   (b) If a value lies within a 95% confidence interval then it will also lie within the 99% confidence interval computed on the same data.

   This is **true**. The 99% confidence interval will be wider than the 95% confidence interval.

   (c) We cannot make a Type II error if the null hypothesis is true.

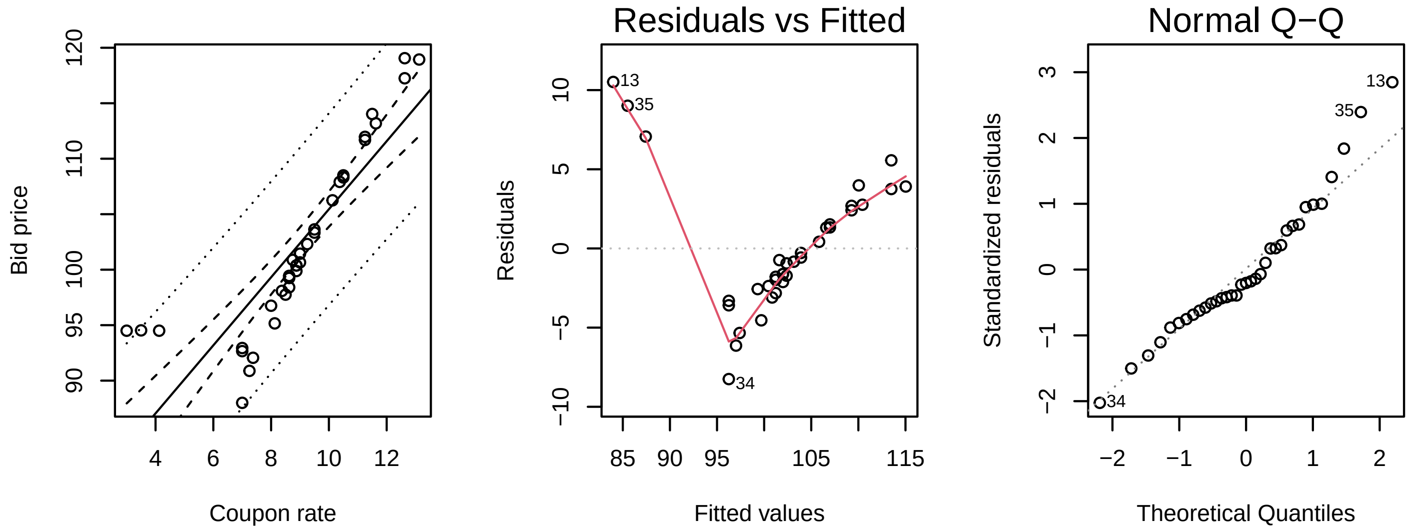
   This is **true**. A Type II error is failing to reject the null hypothesis when the null hypothesis is false.

   (d) If we increase the significance level α of a test of hypotheses, then we increase the probability of making a Type I error.

   This is **true**. The significance level α is maximum allowed probability of making a Type I error.

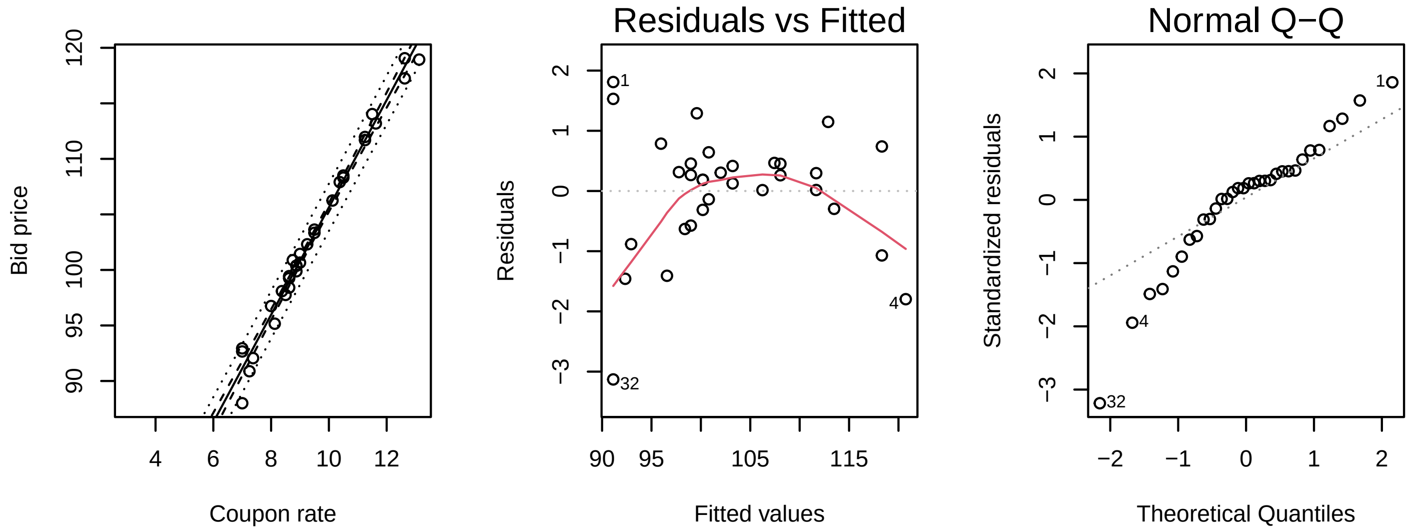
7. Consider two analyses of a data set of bid prices (selling price) versus the coupon rates (the payout to the holder every six months) on November 9, 1988, of US Treasury bonds maturing between 1994 and 1998. The leftmost panel in each analysis shows a scatterplot with the least-squares line overlaid, along with confidence and prediction intervals over the range of the observed data.

Analysis 1:



```
        Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
coupon_rate   3.0661     0.3068   9.994 1.64e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Analysis 2:



```
        Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  57.2932     1.0358   55.31   <2e-16 ***
coupon_rate   4.8338     0.1082   44.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) What accounts for the difference between Analysis 1 and Analysis 2?

> There are three data points which are included in Analysis 1 which are not included in Analysis 2. These data points appear to have had a large influence on the least-squares line.

(b) For both Analysis 1 and Analysis 2, give the estimated increase in the expected bid price due to an increase by one unit in the coupon rate.

> In Analysis 1 it is estimated to be 3.0661, while in Analysis 2 it is estimated to be 4.8338.

(c) Considering Analysis 2, give your conclusion about the null hypothesis H_0: $\beta_1 = 0$.

> We reject the null hypothesis because the *p*-value is very small—less than 2×10^{-16}.

(d) In Analysis 1, what does the residuals versus fitted values plot reveal?

> It shows that there are some extreme outliers in the data set.

(e) In Analysis 2, what does the residuals versus fitted values plot reveal?

> The plot exhibits some fanning in the residual as well as some possibility that the relationship is nonlinear.

(f) Use Analysis 2 to predict the bid price of a US treasury bond which has a coupon rate of 10.

> The predicted bid price is $57.2932 + 4.8338(10) = 105.6303$.

(g) Explain why the confidence and prediction intervals are much narrower in Analysis 2.

> With removal of the outliers, the least-squares line is able to follow the trend apparent in the data, resulting in residuals that are much smaller. These smaller residuals lead to a smaller estimate of the noise variance, which results in much narrower confidence and prediction intervals.