Q6 review key

Stat 301

Summer 2019

(1) *Carbonation*: Corrosion of steel reinforcing bars is the most important durability problem for reinforcing structures. Carbonation of concrete results form a chemical reaction that lowers the pH value by enough to initiate corrosion of the rebar. Data on the carbonation depth (mm) and strength (MPa) for a sample of core specimens was taken from a particular building, and all the regression output is provided. We are interested in modeling the strength.

(a) State the regression model and define its components

$$Y_i = \beta_0 + \beta_1 + \epsilon_i$$

Where:

- y_i: response
- β_0: y-intercept when $x = 0$
- β_1: slope
- x: explanatory variable
- ϵ_i: error term (residual)

(b) Looking at the raw data scatterplot, does it appear as if there is a linear relationship? Positive or negative slope?
There appears to be a linear relationship with a negative slope.

(c) The regression equation is $\hat{y} = 27.18 - 0.298x$. Estimate the strength when the carbonation depth is 8 mm and estimate it again when the depth is 20 mm.

$$\hat{y}|_{x=8} = 27.183 - 0.298(8) = 24.799 \text{ MPa}$$

$$\hat{y}|_{x=20} = 27.183 - 0.298(20) = 21.223 \text{ MPa}$$

(d) Calculate the residuals for both of your estimates in part c. The observed value for 8 mm is 22.8 MPa ($(8, 22.8)$) and for 20 mm is 17.1 MPa ($(20, 17.1)$).

$$e = y - \hat{y}$$

$$e|_{x=8} = 22.8 - 24.799 = -1.999 \text{ since it is negative, this is an overestimate}$$

$$e|_{x=20} = 17.1 - 21.223 = -4.123 \text{ since it is negative, this is an overestimate}$$

(e) Interpret slope and intercept in context of the data. If something does not make sense in context, state it and describe why.
Slope: a one mm increase in the carbonate depth will reduce (because the slope is negative) the strength by 0.296 MPa.

Intercept: when depth is 0 mm, the strength is 27.183 MPa. Even though $x = 0$ is not in our dataset, this could make logical sense in context.

(f) Do a significance test of the slope. State hypotheses, t statistic, *pvalue*, results, and conclusion of the test.

$$H_0 : \beta_1 = 0 \text{ vs. } H_a : \beta_1 \neq 0$$

Test statistic is $t = -7.229$ with $pvalue = 2.01e^{-06} = 2.02 \times 10^{-6} \approx 0 \leq \alpha(0.05)$ so we will reject H_0; the slope is significant.

(g) State, define, and describe R^2 and r (R^2 is on the output and r will require a calculation from the output). R^2, also called the *Coefficient of Determiniation*, is the proportion or percent of the variation in the response that can be explained by the linear relationship. $R^2 = 0.7656 = 75.56\%$ (it is `Multiple R-squared` on the output); 75.56% of the variation in the strength can be accounted for because of the linear relationship between carbonation depth and strength. Correlation is the strength and direction of the *linear* relationship between x and y. Correlation is $r = \sqrt{R^2} = -0.875$; there is a strong, negative linear relationship between carbonation depth and strength.

(h) List assumptions of regression. Are the assumptions of regression met? Briefly explain how each assumption is met or not

1. $E(\epsilon_i) = 0$: the mean of the residuals is 0 (histogram looks a bit odd but is ok so met)

2. $V(\epsilon_i) = \sigma_\epsilon^2$: the variance of the residuals is constant (the same) for all values of y. Also called constant variance, homogeneity of variance (means same variance) (no pattern so met)

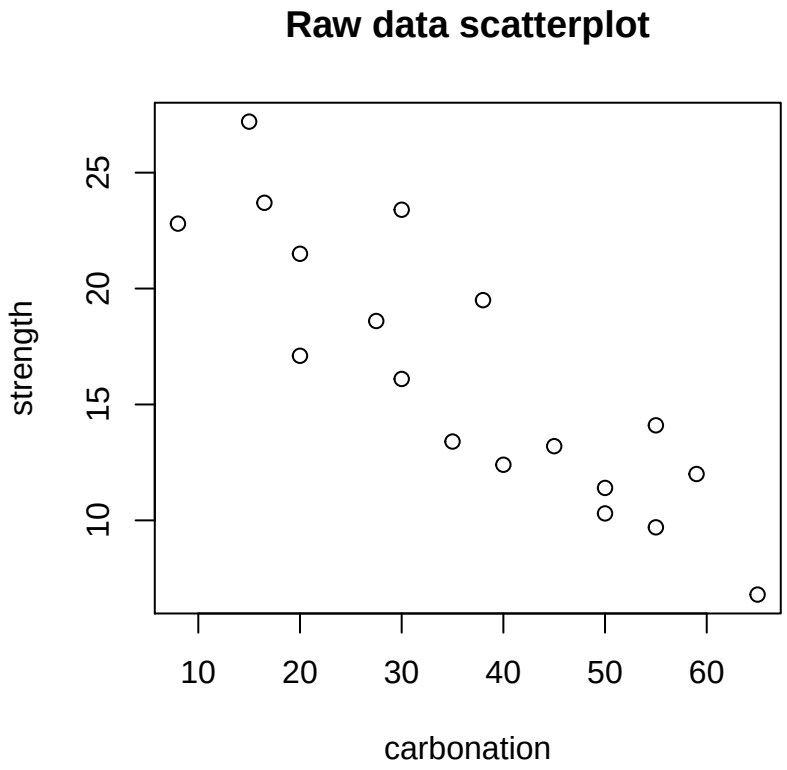
3. $Cov(\epsilon_i, \epsilon_j) = 0$ (independence of residuals – no need to check)

4. $\epsilon_i \sim N(0, \sigma_\epsilon^2)$: Residuals have an approximately normal distribution with mean 0 and homogeneous variance (the qqplot is ok so met)

(i) How is the model? Good, bad, etc.? Give specific evidence (use answers from parts f, g, and h)

Since there is a linear relationship, the slope test was significant (rejection of $H_0 : \beta_1 = 0$), R^2 and r were both good, and the assumptions are met, this is a good model.

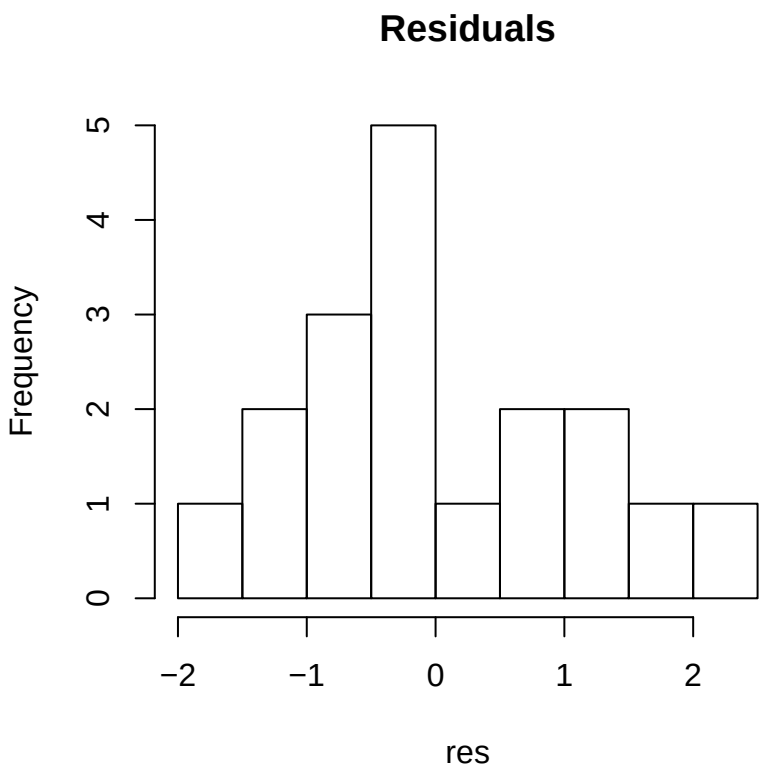
```r
carbonation=c(8,15,16.5,20,20,27.5,30,30,35,38,40,45,50,50,55,55,59,65)
strength=c(22.8,27.2,23.7,17.1,21.5,18.6,16.1,23.4,13.4,19.5,12.4,13.2,11.4,10.3,14.1,9.7,12,6.8)
plot(strength~carbonation,main='Raw data scatterplot')
```



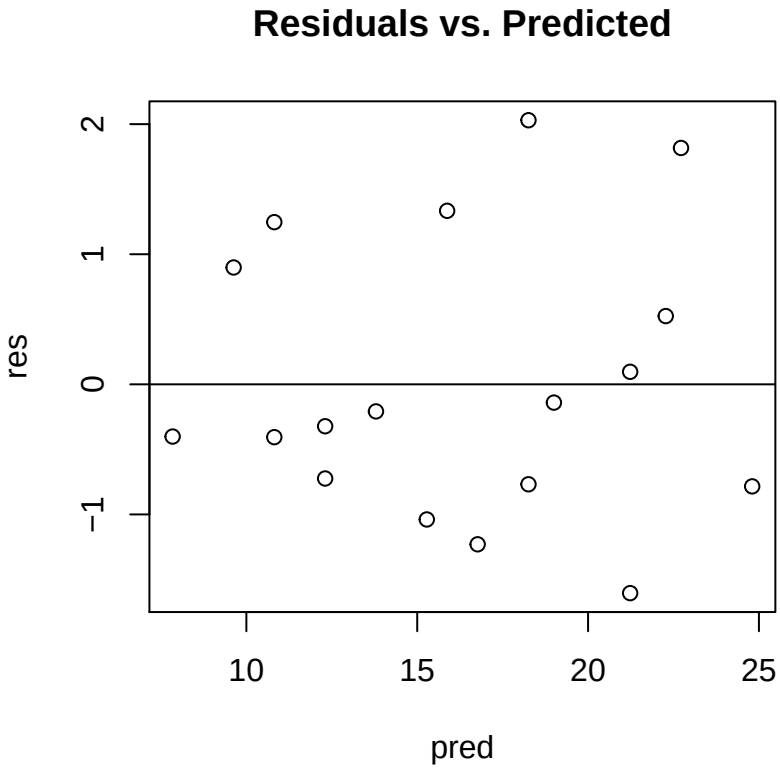

```r
soda=lm(strength~carbonation); summary(soda)
```

```
Call:
lm(formula = strength ~ carbonation)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1317 -2.0043 -0.7488  2.1366  5.1439

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.18294    1.65135  16.461 1.88e-11 ***
carbonation -0.29756    0.04116  -7.229 2.01e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.864 on 16 degrees of freedom
Multiple R-squared:  0.7656,    Adjusted R-squared:  0.7509
F-statistic: 52.25 on 1 and 16 DF,  p-value: 2.013e-06
```

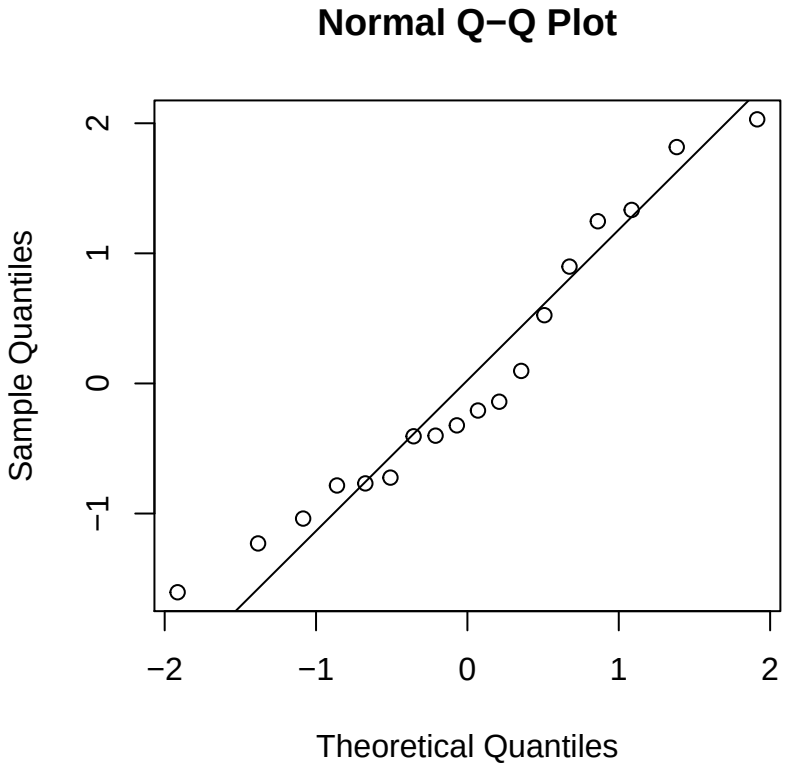
```r
res=rstudent(soda); pred=fitted(soda)
hist(res,main='Residuals')
```




```
plot(pred,res,main='Residuals vs. Predicted'); abline(0,0)
```



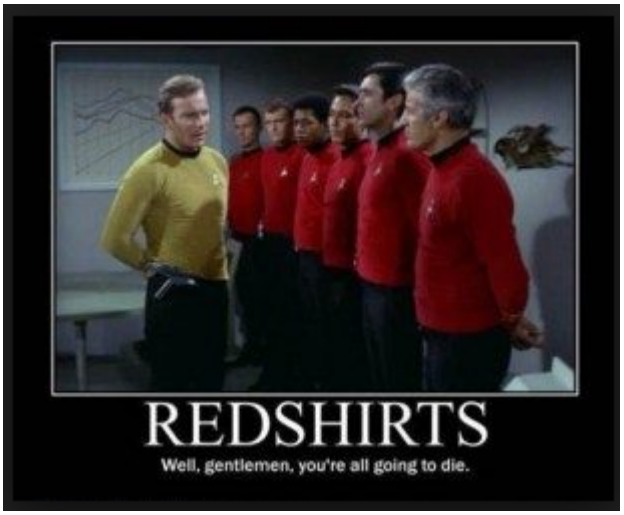

```
qqnorm(res); qqline(res)
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
−1
0
1
2

(2) *Red Shirt of Dooooom*: In Star Trek fandom, there is a running joke that characters on the show who wear a red shirt are doomed[1]; just another statistic. Shirt colors can be only blue, gold, or red; fatalities can be only dead or alive.

(a) Is there sufficient evidence determine whether there is an association between shirt color and deaths?

(b) State the kind of error that could have been made. *Describe in context*




```
        Shirt.Colour
Survival Blue Red Gold Total
   Alive  129 215   46   390
   Dead     7  24    9    40
   Total  136 239   55   430
```

Star Trek survival by shirt colour

Null Hypothesis
H_0 : Shirt colour and survival on Star Trek are independent

Alternative hypothesis
H_a : H_0 is not true (Shirt colour and survival on Star Trek are dependent)

Expected values

$$E_{ij} = \frac{n_i n_j}{n} = \frac{(rtotal)(ctotal)}{grandtotal}$$

$E_{11} = (390 * 136/430) = 123.35$
$E_{12} = (390 * 239/430) = 216.77$
$E_{13} = (390 * 55/430) = 49.88$
$E_{21} = (40 * 136/430) = 12.65$
$E_{22} = (40 * 239/430) = 22.23$
$E_{23} = (40 * 55/430) = 5.12$

[1]Honestly, depending on which season or episodes the sample is taken from, the results of the test can vary. When you use the entire population (all *TV episodes* of the origianl Star Trek; no movies or other Star Trek series), the conculsion is that shirt colour and survival are independent (besides, Chief Engineer Scotty wore red... he was around for a while and *did* go on the ocassional away mission).

Here we can check to see all $E_{ij} \geq 5$

Test Statistic

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum \frac{(observed - expected)^2}{expected}$$

$= \frac{(129-123.35)^2}{123.35} + \frac{(215-216.77)^2}{216.77} + \frac{(46-49.88)^2}{49.88} + \frac{(7-12.65)^2}{12.65} + \frac{(24-22.23)^2}{216.77} + \frac{(9-5.12)^2}{5.12} = 6.189$

$df = (r-1)(c-1)$ where $r, c = (2, 3)$ so $df = (2-1)(3-1) = 2$

Rejection Region
1. *Critical Value approach*: Reject H_0 if $\chi^2_{calc} \geq \chi^2_{\alpha,df}$ where $\chi^2_{\alpha,df} = \chi^2_{.05,2} = 5.991$
2. *pvalue approach*: Reject H_0 if $pvalue \leq \alpha$

Results
We are doing the critical value approach: $\chi^2_{.05,2} = 5.991$. $6.189 \geq 5.991$ so we will reject H_0.

Conclusion (in context)
We rejected H_0 so that tells us that the dreaded red shirt does mean you are less likely to survive an episode of Star Trek (survival is dependent on shirt colour).

Error
We rejected H_0 so a type I error could have been made. We think that survival depends on shirt colour when shirt colour makes no difference in survival.

(3) *Here be dragons*: An analysis of dragon reserve accident data was made to determine if there is a relationship between the type of accident (fatal or non-fatal) and the location of the dragon reserve (Romania, Canada, Austrailia). Is there sufficent evidence that more fatal injuries happen at one or two specific reserves? The data for 346 accidents are shown in the accompanying table.
   (a) Is there sufficient evidence determine if injury type is the same at all dragon reserve locations?
   (b) State the kind of error that could have been made. *Describe in context*

	Location			
Survival	Romania	Canada	Australia	Total
Fatal	128	63	46	237
Non Fatal	67	26	16	109
Total	195	89	62	346

Dragon Reserve Injuries by Location

Null Hypothesis
H_0 : Fatal and non-fatal injuries are the same in all dragon reserves

Alternative hypothesis
H_a : H_0 is not true (Fatal and non-fatal injuries are different across dragon reserves)

Expected values

$$E_{ij} = \frac{n_i n_j}{n} = \frac{(rtotal)(ctotal)}{grandtotal}$$

$E_{11} = (237 * 195/346) = 133.57$
$E_{12} = (237 * 89/346) = 60.96$
$E_{13} = (237 * 62/346) = 42.47$
$E_{21} = (109 * 195/346) = 61.43$
$E_{22} = (109 * 89/346) = 28.04$
$E_{23} = (109 * 55/346) = 19.53$

Here we can check to see all $E_{ij} \geq 5$

Test Statistic

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum \frac{(observed - expected)^2}{expected}$$

$= \frac{(128-133.57)^2}{133.57} + \frac{(63-60.96)^2}{60.96} + \frac{(46-42.47)^2}{42.47} + \frac{(67-61.43)^2}{61.43} + \frac{(26-28.04)^2}{28.04} + \frac{(16-19.53)^2}{19.53} = 1.886$
$df = (r-1)(c-1)$ where $r, c = (2, 3)$ so $df = (2-1)(3-1) = 2$

Rejection Region

Reject H_0 if $\chi^2_{calc} \geq \chi^2_{\alpha,df}$ where $\chi^2_{\alpha,df} = \chi^2_{.05,2} = 5.991$

Results
$\chi^2_{.05,2} = 5.991$. $1.886 \ngeq 5.991$ so we cannot reject H_0.

Conclusion (in context)
We did not reject H_0, indicating that the injury type is the same at all dragon reserves.

Error
We did not reject H_0 so a type II error could have been made. We think that injury type is the same at all dragon reserves, when there are differences in injuries by reserve.

(4) *Book Mediums*: (not the psychic kind of medium) A professor of an introductory college class uses an open-source textbook for the class. Of interest is the proportions of students that will either purchase a hard copy, print the book online, or just use the downloaded PDF format to read on a device. From earlier semesters, 60% bought a hard copy of the book, 25% printed it online, and 15% used a downloaded PDF format on their devices. At the end of the semester, the professor asks the students to complete a survey and indicate what format of the book they used. Of the 126 students, 71 bought a hard copy, 30 printed it, and 25 downloaded PDF to use.
   (a) Is there evidence that the students used similar mediums for the book?
   (b) State the kind of error that could have been made. *Describe in context*

```
       type counts probs
1 Hard copy     71  0.60
2   Printed     30  0.25
3       PDF     25  0.15
```

Null Hypothesis
H_0 : Students use of books is 60% hard copy, 25% printed, 15% PDF (or $H_0 : p_1 = 0.6,\ p_2 = 0.25,\ p_3 = 0.15$)

Alternative hypothesis
H_a : H_0 is not true (students use of books are not the estimated percents as listed above)

Expected values

$$E_i = np_i$$

$E_1 = 126(0.6) = 75.6$
$E_2 = 126(0.25) = 31.5$
$E_3 = 126(0.15) = 18.9$

Here we can check to see all $E_{ij} \geq 5$

Test Statistic

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \sum \frac{(observed - expected)^2}{expected}$$

$= \frac{(71-75.6)^2}{75.6} + \frac{(30-31.5)^2}{31.5} + \frac{(25-18.9)^2}{18.9} = 2.32$

$df = k - 1$ where $k = 3$ so $df = 3 - 1 = 2$

Rejection Region

Reject H_0 if $\chi^2_{calc} \geq \chi^2_{\alpha,df}$ where $\chi^2_{\alpha,df} = \chi^2_{.05,2} = 5.991$

Results
$\chi^2_{.05,2} = 5.991$. $2.32 \ngeq 5.991$ so we cannot reject H_0.

Conclusion (in context)
We did not reject H_0, indicating that the students are using the different book mediums as expected (similar to other semesters).

Error
We did not reject H_0 so a type II error could have been made. We think that students are using the different mediums of books as expected but they are not.