STAT 408 Homework 2 Solution

1.

a. y = 10 + 0.56 * 7 = 13.92

b. residual = 17 – (10 + 0.56 * 7) = 3.08

c. response will increase by one β_1 = 0.56

d. Not necessary, because the model includes error term. The test score will be slightly different from 17.

2.

```
teengamb$sex <- factor(teengamb$sex)

levels(teengamb$sex) <- c("Male", "Female")

model <- lm(gamble~sex+status+income+verbal, data = teengamb)

summary(model)
```

```
Call:
lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)

Residuals:
    Min      1Q  Median      3Q     Max
-51.082 -11.320  -1.451   9.452  94.252

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.55565   17.19680   1.312   0.1968
sexFemale   -22.11833    8.21111  -2.694   0.0101 *
status        0.05223    0.28111   0.186   0.8535
income        4.96198    1.02539   4.839 1.79e-05 ***
verbal       -2.95949    2.17215  -1.362   0.1803
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.69 on 42 degrees of freedom
Multiple R-squared:  0.5267,    Adjusted R-squared:  0.4816
F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

b.

R-square measures the percentage of variation in the response explained by predictors, which is 0.5267, the “Multiple R-square” in previous output.

c.

which.max(model$residuals)

The 24th observation has the largest residual 94.25222.

1	2	3	4	5	6	7	8	9	10
10.6507430	9.3711318	5.4630298	-17.4957487	29.5194692	-2.9846919	-7.0242994	-12.3060734	6.8496267	-10.3329505
11	12	13	14	15	16	17	18	19	20
1.5934936	-3.0958161	0.1172839	9.5331344	2.8488167	17.2107726	-25.2627227	-27.7998544	13.1446553	-15.9510624
21	22	23	24	25	26	27	28	29	30
-16.0041386	-9.5801478	-27.2711657	94.2522174	0.6993361	-9.1670510	-25.8747696	-8.7455549	-6.8803097	-19.8090866
31	32	33	34	35	36	37	38	39	40
10.8793766	15.0599340	11.7462296	-3.5932770	-14.4016736	45.6051264	20.5472529	11.2429290	-51.0824078	8.8669438
41	42	43	44	45	46	47			
-1.4513921	-3.8361619	-4.3831786	-14.8940753	5.4506347	1.4092321	7.1662399			

d.

cor(model$residuals,model$fitted.values)

The correlation is -1.070659e-16, almost zero. The page 14 in lecture slides shows that residual should be orthogonal/independent to fitted response. Therefore, the correlation should be zero.

1	2	3	4	5	6	7	8	9	10
-10.6507430	-9.3711318	-5.4630298	24.7957487	-9.9194692	3.0846919	8.4742994	18.9060734	-5.1496267	10.4329505
11	12	13	14	15	16	17	18	19	20
-1.4934936	8.4958161	1.0827161	-5.9331344	-0.4488167	-13.8107726	25.3627227	36.1998544	-1.1446553	15.9510624
21	22	23	24	25	26	27	28	29	30
17.0041386	10.7801478	27.3711657	61.7477826	37.8006639	11.2670510	40.3747696	11.7455549	7.4803097	29.4090866
31	32	33	34	35	36	37	38	39	40
77.1206234	38.1400660	78.2537704	6.5932770	28.5016736	24.3948736	17.9527471	45.9570710	57.0824078	16.1330562
41	42	43	44	45	46	47			
8.3513921	73.5361619	17.6831786	15.4940753	32.5493653	12.9907679	12.0337601			

e.

cor(model$residuals,teengamb$income)

The correlation is -7.242382e-17, almost zero. The page 14 in lecture slides shows that residual should be orthogonal/independent to the plane spanned by X. Therefore, the correlation between residual and any predictors should be zero.

f.

If all other predictors held constant, moving from male to female will decrease gambling by 22.11833.

3.

a.

model <- lm(lpsa ~ lcavol , data = prostate)

summary(model)

deviance(model)

The RSS is 58.91476 and R-square is 0.5394

b.

model <- lm(lpsa ~ . , data = prostate)

summary(model)

deviance(model)

The RSS is 44.16302 and R-square is 0.6548

c.

RSS is lower and R-square is higher in the second model. With more predictors, the model will explain more variance of the response and thus generate less residuals. This is for sure regardless of the significance of extra predictors in the second model.

d.

x <- model.matrix(~ lweight + age + lbph + svi + lcp + gleason + pgg45 + lcavol, prostate)

y <- prostate$lpsa

xtxi <- solve(t(x) %*% x)

xtxi %*% t(x) %*% y

```
                    [,1]
(Intercept)  0.669336698
lweight      0.454467424
age         -0.019637176
lbph         0.107054031
svi          0.766157326
lcp         -0.105474263
gleason      0.045141598
pgg45        0.004525231
lcavol       0.587021826
```

The manual result is almost identical to the lm function.

4.

a.

model <- lm(taste~.,cheddar)

summary(model)

```
Coefficients:
            Estimate
(Intercept) -28.8768
Acetic        0.3277
H2S           3.9118
Lactic       19.6705
```

b.

cor(model$fitted.values, cheddar$taste)

The correlation is 0.8073256. The fitted response is highly correlated to the true response. It shows a relatively good model fit and a small residual sum of squares.

c.

The intercept means that if all three chemical contents are zeros, the average taste score is -28.8768. In the dataset, the minimum score is positive. Therefore, the negative score with such large scale would be counter intuitive.

5.

a.

This is a simulation study. The code first randomly generated 100 numbers from a uniform distribution U[0, 10]. Then it defined a true linear model by using those numbers as one predictor:

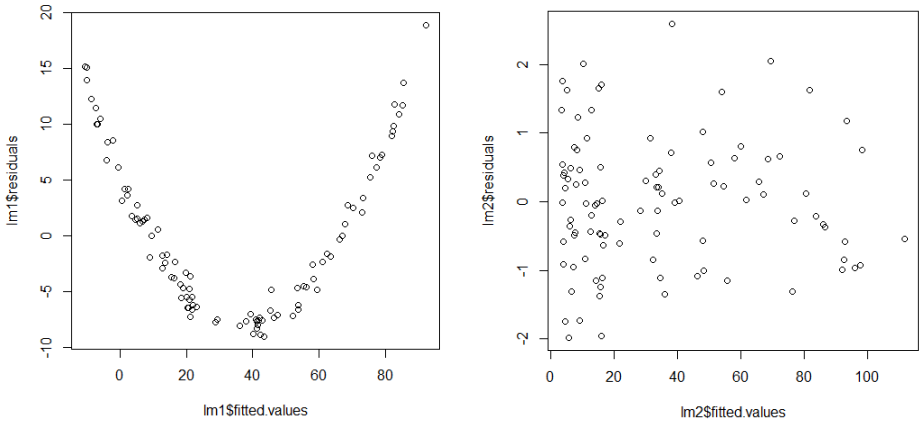
$$y = 3 + x + x^2 + \varepsilon$$

where error $\varepsilon \sim N(0, 1)$. Finally, it fits two linear models on the simulated date, first with one predictor x, and second with two predictors x and x^2.

b.

plot(lm1$fitted.values, lm1$residuals)

plot(lm2$fitted.values, lm2$residuals)



The residual of the first model shows a quadratic pattern, while the second one does not have strong pattern and is close to random noise.

c.

The second one is better because it correctly specified the model form. The first one missed the x^2. Also the residual of second model is much closer to a normal distribution $N(0,1)$ or random noise.