Stat 500 - Homework 3 (Solutions)

Part A.

1. We fit a linear model with `total` sat score as response and `takers`, `ratio`, and `salary` as predictors. The R-squared is 0.8239, i.e. the three predictors explain about 82.39% of the variation in the response variable.

However, this information alone is not sufficient to decide whether the model is a good fit to the data. Always visualize the data, residuals, and fitted values to check for nonlinear relationships between response and predictors, and to see whether the assumptions necessary for hypothesis testing are met. Here, we skip over those steps to keep the solution concise. (But if you did go through those steps, you would see that all assumptions are met reasonably well!)

```
> library(faraway)
> data(sat)
> names(sat)
[1] "expend" "ratio"  "salary" "takers" "verbal" "math"   "total"
> fit <- lm(total ~ takers + ratio + salary, data=sat)
> summary(fit)

Call:
lm(formula = total ~ takers + ratio + salary, data = sat)

Residuals:
    Min      1Q  Median      3Q     Max
-89.244 -21.485  -0.798  17.685  68.262

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1057.8982    44.3287  23.865   <2e-16 ***
takers        -2.9134     0.2282 -12.764   <2e-16 ***
ratio         -4.6394     2.1215  -2.187   0.0339 *
salary         2.5525     1.0045   2.541   0.0145 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.41 on 46 degrees of freedom
Multiple R-squared:  0.8239,Adjusted R-squared:  0.8124
F-statistic: 71.72 on 3 and 46 DF,  p-value: < 2.2e-16
```

2. $H_0 : \beta_3 \leq 0$ versus $H_1 : \beta_3 > 0$. The test statistic for this test is $t = \frac{\hat{\beta}_3 - 0}{\widehat{s.e.}(\hat{\beta}_3)} \sim t_{46}$.[1] From the R output we have that $t = 2.541$ and that the p-value for the two-sided test $H_0 : \beta_3 = 0$ versus $H_1 : \beta_3 \neq 0$ is $P(|t_{46}| > |2.541|) = 0.0145$. Therefore, the p-value for our one-sided hypothesis test is $P(t_{46} > 2.541) = 0.0145/2 = 0.00725$. Thus, at a significance level of $\alpha = 0.01$ we reject the null hypothesis that β_3 is non-positive.

[1] Note that this is actually the test statistic associated with null hypothesis $\beta_3 = 0$. However, if this test statistic leads us to reject the null hypothesis $\beta_3 = 0$, then we also reject that $\beta_3 = x$ for any $x < 0$. Why? Think about what p-values mean!

3. $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$. The test statistic for this test is $t = \frac{\hat{\beta}_2 - 0}{\widehat{s.e.}(\hat{\beta}_2)} \sim t_{46}$. From the R output we have $t = -2.187$ and a p-value of $P(|t_{46}| > |-2.187|) = 0.0339$. Thus, at a significance level of $\alpha = 0.01$ we fail to reject the null hypothesis that β_2 does not have an effect on the SAT scores in the full model. What other test could you use to answer this question?
4. $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ versus H_1 : "at least one regression coefficient is not 0". This can also be phrased as testing the reduced model (which does not contain any predictors) against the full model (which includes all predictors). The test statistic for this test is

$$F = \frac{(RSS_{reduced} - RSS_{full})/(49 - 46)}{RSS_{full}/46} \sim F_{3,46}.$$

From the R output we have $F = 71.72$ with associated p-value equal to $2.2 \times 10^{-16} \approx 0$. Hence, for any significance level $\alpha > 0$ we reject the null hypothesis that no predictor is relevant to explain the SAT scores.
5. The CI's are given below. Note that the 95% CI does not contain 0, whereas the 99% CI does contain 0. Hence, we conclude that the p-value lies in the interval $(0.01, 0.5)$.

```
> confint(fit, level=0.95)["salary",]
    2.5 %    97.5 %
0.5304797 4.5744605
> confint(fit, level=0.99)["salary",]
    0.5 %    99.5 %
-0.146684  5.251624
```

6. We use the code from lecture 3 to produce the joint confidence region for parameters associated with `ratio` and `salary`:

```
library(ellipse)
# Plot the confidence region
plot(ellipse(fit, c('ratio', 'salary')), type="l")
# Add the estimates to the plot
points(fit$coef['ratio'], fit$coef['salary'],pch=18)
# Add the origin to the plot
points(0, 0, pch=19, col="red")
# Add the confidence intervals
conf <- confint(fit, level=0.95)
abline(v=conf['ratio',], lty=2)
abline(h=conf['salary',], lty=2)
```

Note that the origin lies outside the 95% joint confidence region. Therefore, if we were to test $H_0 : \beta_2 = \beta_3 = 0$ versus H_a : "at least one of the two coefficients β_2 and β_3 is not zero", we would reject H_0 at a 5% significance level.
7. We add `expend` to the linear model:

```
> fit2 <- lm(total ~ takers + ratio + salary + expend, data=sat)
> summary(fit2)

Call:
lm(formula = total ~ takers + ratio + salary + expend, data = sat)
```

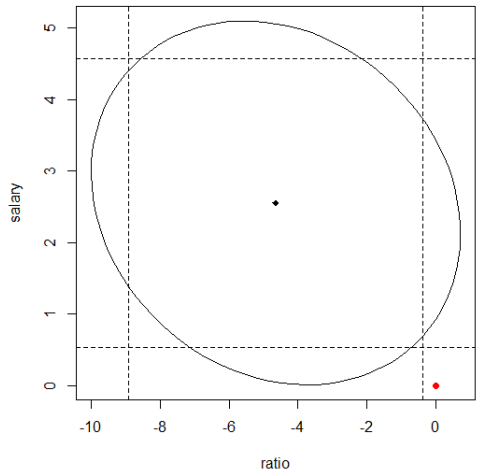

Figure 1: 95% Confidence Region for the parameters associated with `ratio` and `salary`.

```
Residuals:
    Min      1Q  Median      3Q     Max
-90.531 -20.855  -1.746  15.979  66.571

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
takers        -2.9045     0.2313 -12.559 2.61e-16 ***
ratio         -3.6242     3.2154  -1.127    0.266
salary         1.6379     2.3872   0.686    0.496
expend         4.4626    10.5465   0.423    0.674
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.7 on 45 degrees of freedom
Multiple R-squared:  0.8246,Adjusted R-squared:  0.809
F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

The variables `ratio`, `salary`, and `expend` are all insignificant at any significance level. Furthermore, the adjusted R-squared of `fit2` is less than the adjusted R-squared of `fit1`. Hence, adding the additional regressor does not improve the goodness of fit.

8. This is a test on nested models with the reduced model containing only predictor `takers` and the alternative model being the full model with all four predictors `takers`, `ratio`, `salary`, and `expend`. We use an F-test to decide which model is better:

```
> fit3 <- lm(total ~ takers, data=sat)
> anova(fit3, fit2)
Analysis of Variance Table

Model 1: total ~ takers
```

```
Model 2: total ~ takers + ratio + salary + expend
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     48 58433
2     45 48124  3     10309 3.2133 0.03165 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the R output we see that at a 5% significance level we fail to reject that null hypothesis that the reduced model containing only predictor `takers` is the better model.

Part B.

There are many ways to do this problem. Here is one possible solution. We first check the structure of the relationship between the predictors and the response:

```
> data(teengamb)
> pairs(teengamb,col=as.numeric(teengamb$sex)+2)
```

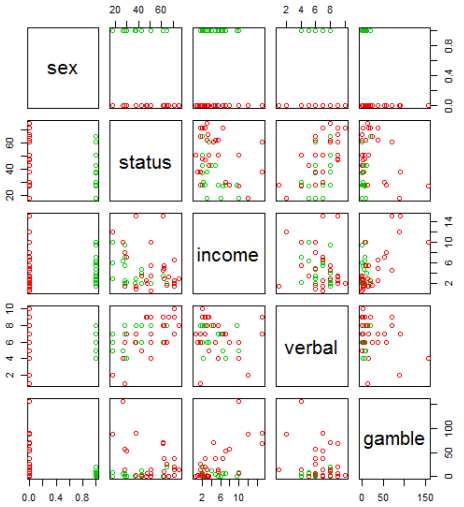

Figure 2: Scatterplot of `teengamb` data set.

The scatterplot in Figure 2 shows that only variables `sex` and `income` are significantly correlated with variable `gamble`. Furthermore, the relationship seems to be linear. We formally test this observation with an F-test between a linear model containing `sex`, `income`, `status`, and `verbal` as predictors and a reduced linear model containing only `sex` and `income` as predictors.

```
> # Candidate models
> fit1 <- lm(gamble ~ sex + status + income + verbal, data=teengamb)
> #summary(fit1)
> fit2 <- lm(gamble ~ sex + income, data=teengamb)
> #summary(fit2)
> anova(fit2, fit1)
Analysis of Variance Table
```

```
Model 1: gamble ~ sex + income
Model 2: gamble ~ sex + status + income + verbal
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     44 22781
2     42 21624  2    1157.5 1.1242 0.3345
```

Indeed, at any reasonable significance level we fail to reject the null hypothesis that the reduced model is the better model. Hence, from now on we work with the reduced model `fit2`.

Next, we check whether the errors are homoscedastic and approximately normally distributed.

```
> # Homoscedasticity
> plot(fit2$fitted, fit2$res)
> abline(h=0)
> # Normality of errors
> qqnorm(fit2$res, ylab="Residuals")
> qqline(fit2$res)
> hist(fit2$res, xlab="Residuals")
```

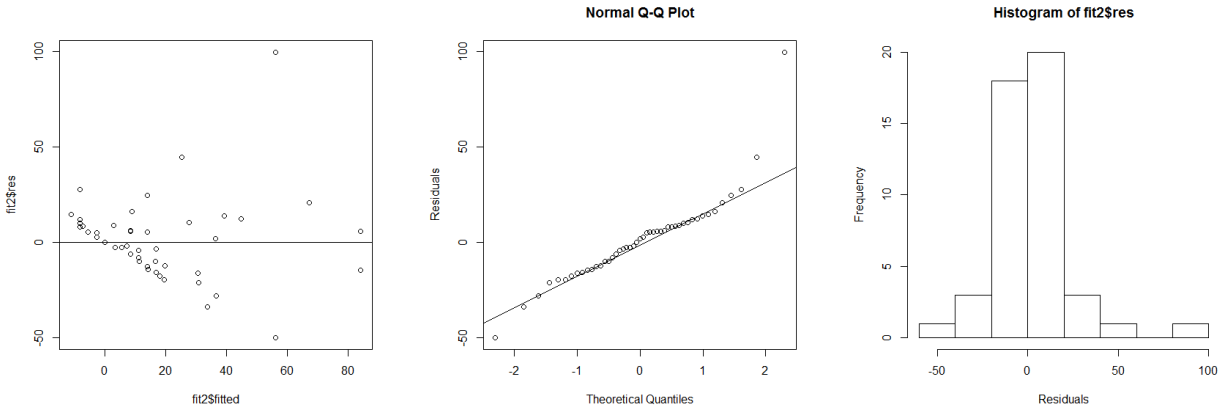

Figure 3: (a) Fitted values versus residuals, (b) QQ-Plot of residuals, and (c) Histogram of residuals.

Clearly, the magnitude of the residuals grows with the fitted values. Hence, the errors are not homoscedastic. Moreover, the shape of the QQ-plot suggests that the errors have heavier lower and upper tails than Gaussian random variables. The histogram is almost symmetric around zero, the longer right tail might be due to an outlier (something that we will examine below). To stabilize the variance we follow the hint and take the square root of the response.

```
> # Transformed Response
> fit3 <- lm(sqrt(gamble) ~ sex + income, data=teengamb)
> # Homoscedasticity
> plot(fit3$fitted, fit3$res)
> abline(h=0)
> # Normality of errors
> qqnorm(fit3$res, ylab="Residuals")
> qqline(fit3$res)
> hist(fit3$res, xlab="Residuals")
```

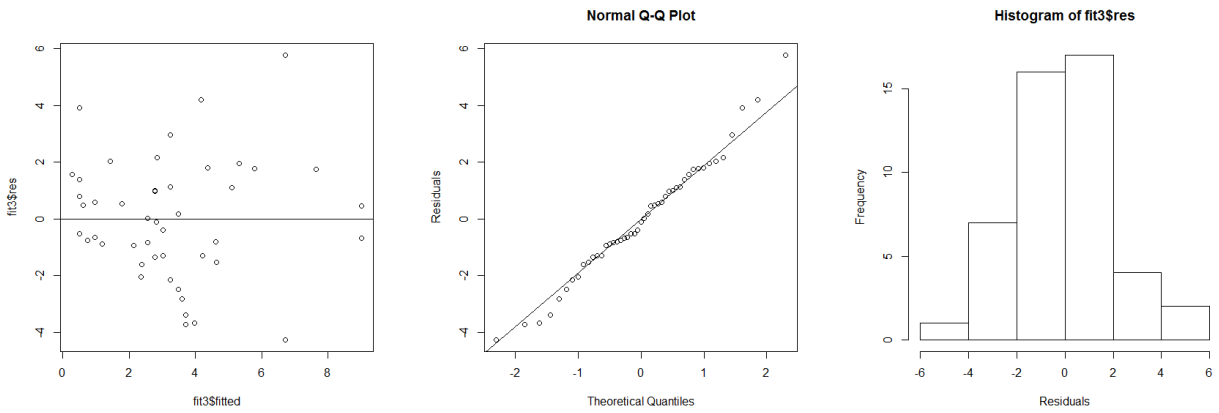

Figure 4: (a) Fitted values versus residuals (transformed response), (b) QQ-Plot of residuals (transformed response), and (c) Histogram of residuals (transformed response).

From Figure 4 we infer that taking the square root of the response yields stable, homoscedastic variances that are approximately Gaussian. Therefore, we keep working with model `fit3`.

Finally, we check for outliers, large leverage, and influential points.

```
> # Compute studentized residuals
> fit3.s <- summary(fit3)
> sigma.s <- fit3.s$sig
> hat.s <- lm.influence(fit3)$hat
> stud.res <- fit3$residuals/(sigma.s * sqrt(1-hat.s))
> plot(stud.res, fit3$residuals, xlab="Studentized residuals", ylab="Raw residuals")
> # Half-normal plot for leverages
> halfnorm(lm.influence(fit3)$hat, nlab = 2, ylab="Leverages")
```

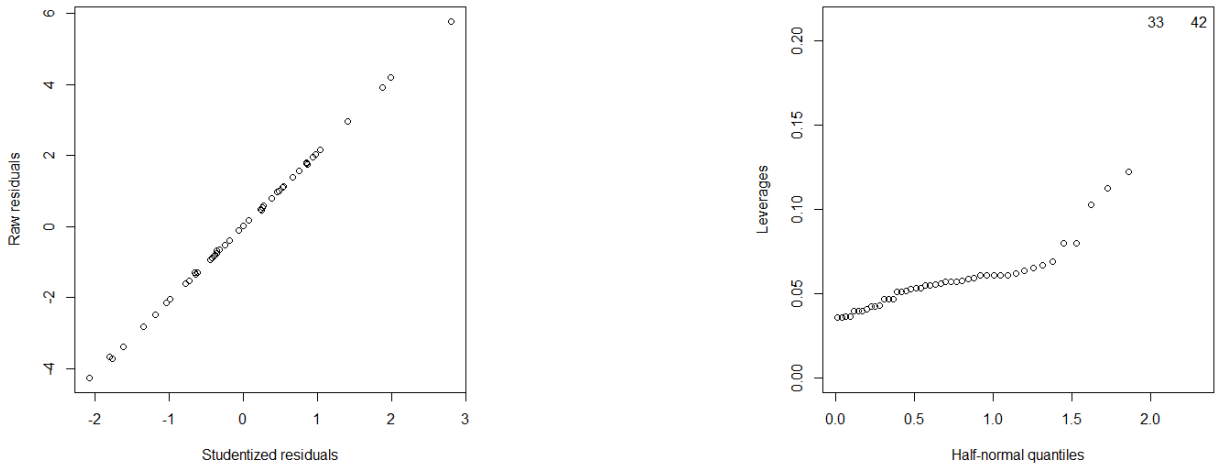

Figure 5: (a) Internally studentized residuals (transformed response), (b) half-normal plot (transformed response).

Figure 5(a) suggests that there are no outliers in the data set, Figure 5(b) suggests that there are two observations with high leverage, observations 33 and 42. Those two points could be influential points. We have re-run our analysis without those points; however, our findings did not change significantly. Therefore, we do not report them here.

Part C.
Let $(y_1, x_1), \ldots, (y_n, x_n)$ be a sample of pairs of response variable y and predictors $x = [A, B]'$. Write $X = [x_1, \ldots, x_n]'$ for the $n \times 2$-dimensional matrix with rows (!) $x_1', \ldots x_n$. By assumption on A and B, and the law of large numbers

$$\frac{1}{n} X'X \approx \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} =: \Sigma,$$

where ρ is the correlation coefficient between A and B. Note that Σ has eigenvalues $1+\rho$ and $1-\rho$ with corresponding normalized eigenvectors $e_1 = \frac{1}{\sqrt{2}}(1,1)'$ and $e_2 = \frac{1}{\sqrt{2}}(1,-1)'$. WLOG assume that $\hat{\beta} = (\hat{\beta}_A, \hat{\beta}_B)' = 0$. Then, the $(1-\alpha)$-confidence region for the estimate $\hat{\beta} = 0$ is the set of all β's satisfying

$$\frac{1}{2\hat{\sigma}^2} \beta' X' X \beta \leq c_\alpha \tag{1}$$

for an appropriate value of $c_\alpha > 0$. Whence, the confidence region can be thought of as an approximate level set of the quadratic form

$$Q(u) = u' \Sigma u.$$

We know that the shape of level sets of quadratic forms is determined by their eigenvalues and eigenvectors. In particular, for any $w = (w_1, w_2) \in \text{span}(e_1)$ and $v = (v_1, v_2) \in \text{span}(e_2)$ we have

$$Q(w) = w' \Sigma w = 2(1+\rho) w_1^2,$$
$$Q(v) = v' \Sigma v = 2(1-\rho) v_2^2.$$

Now, observe that $w \in \text{span}(e_1)$ and $v \in \text{span}(e_2)$ lie on the boundary of the level set corresponding to c_α if and only if $|w_1| = |w_2| = \sqrt{\frac{c_\alpha}{2(1+\rho)}}$ and $|v_1| = |v_2| = \sqrt{\frac{c_\alpha}{2(1-\rho)}}$. Therefore, if $\rho > 0$, then $|w_1| < |v_1|$. Thus, the ellipse defined by 1 is compressed in the direction of eigenvector e_1 and stretched out in direction of eigenvector e_2. This results in the "leaning to the left" effect. See Figure 6.

Similarly, we can conclude that if $\rho < 0$, the ellipse is stretched out in direction of e_1 but compressed in direction e_2. This yields the "leaning to the right" effect. Finally, if $\rho = 0$ then the ellipse is a circle with radius $\sqrt{\frac{c_\alpha}{2}}$; there is no stretching or compressing in any direction.

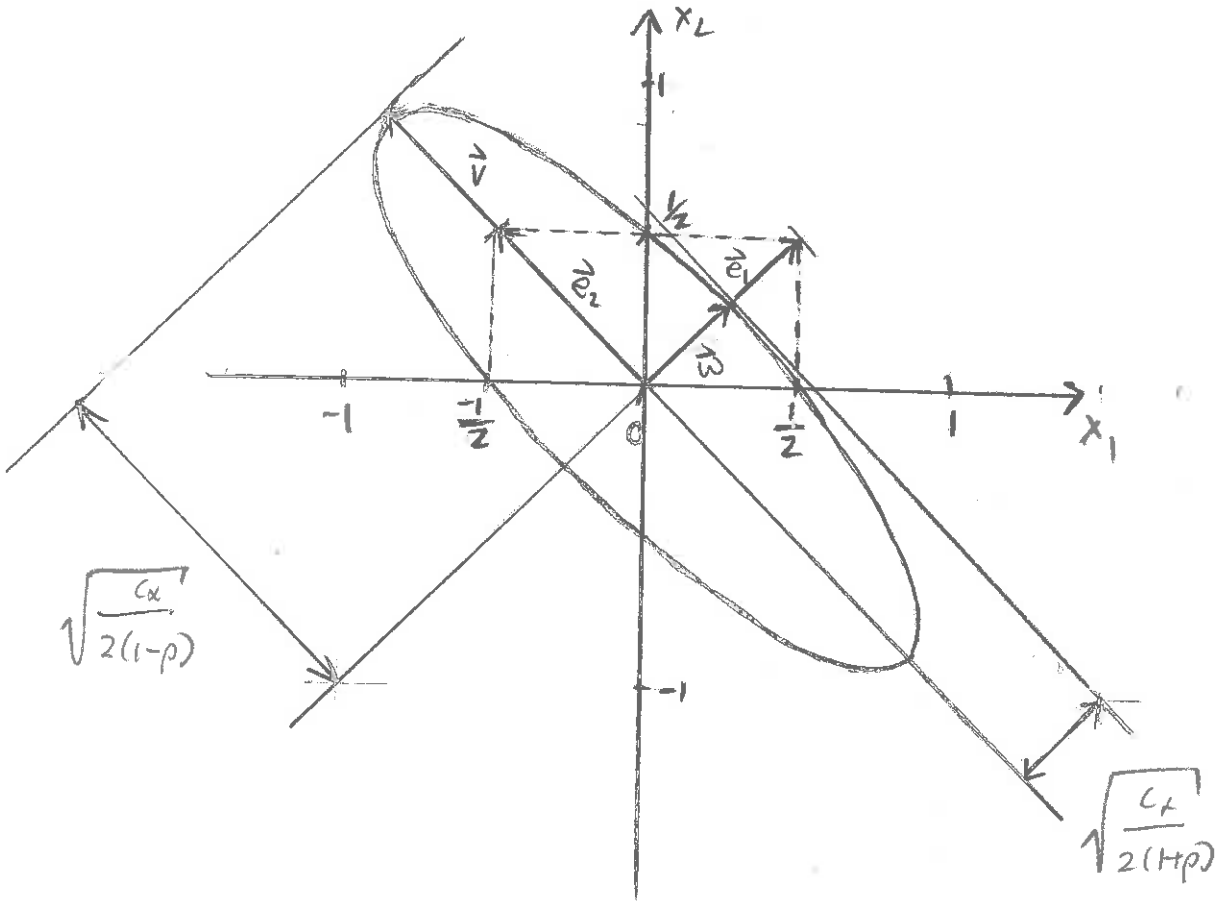

Figure 6: Level set of the quadratic form $Q(u) = u'\Sigma u$ for some $c_\alpha > 0$ and $\rho > 0$.