Homework 03 Solutions

2022-09-15

36-617: Applied Linear Models
Fall 2022
Solutions

```
library(arm)  ## includes lme4, MASS, Matrix
library(ggplot2); theme_set(theme_bw())
library(gridExtra) ## to arrange ggplots...
library(GGally) ## for ggpairs...
```

Problem 1: Sheather, Ch 3 (yes!), pp. 109 ff., #5. % cars04.csv

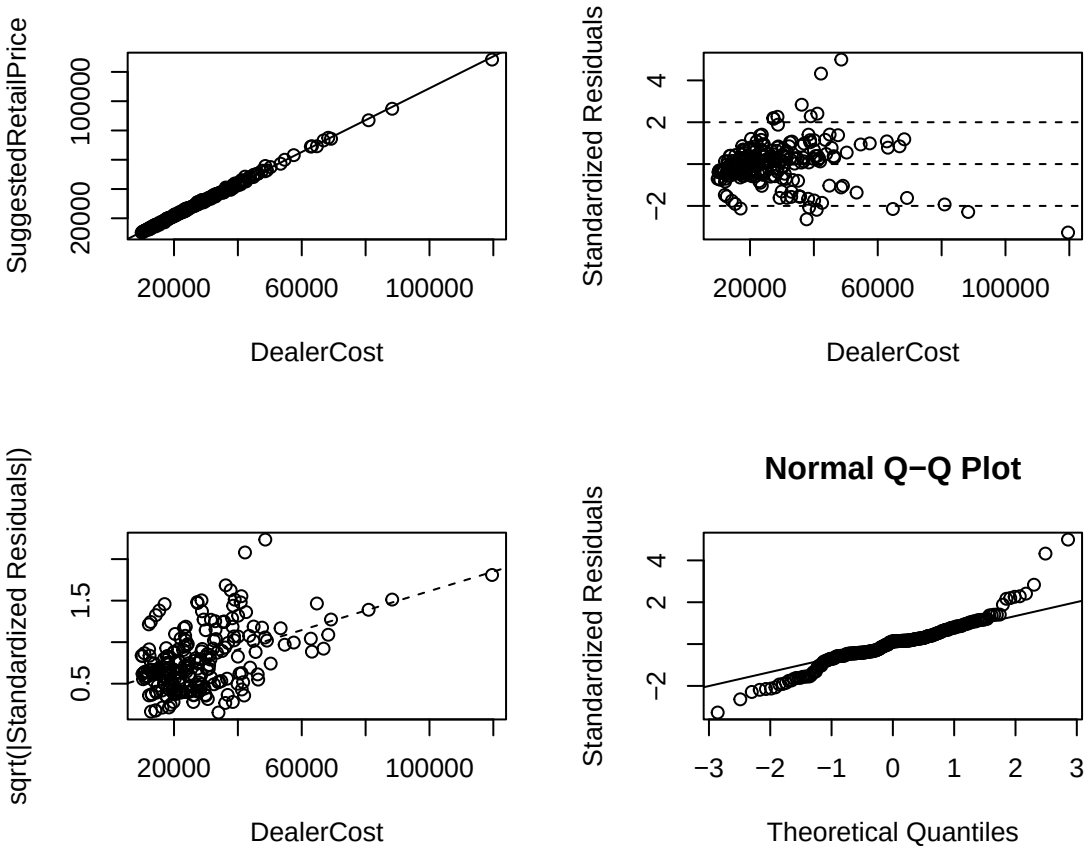
```
cars04 <- read.csv("cars04.csv",header=T)
lm.3.10 <- lm(SuggestedRetailPrice ~ DealerCost,data=cars04)
summary(lm.3.10)
```

```
##
## Call:
## lm(formula = SuggestedRetailPrice ~ DealerCost, data = cars04)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1743.52  -262.59    74.92   265.98  2912.72
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -61.904248  81.801381  -0.757     0.45
## DealerCost    1.088841   0.002638 412.768   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 587 on 232 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9986
## F-statistic: 1.704e+05 on 1 and 232 DF,  p-value: < 2.2e-16
```

Problem 1(a)

Based on the above output for model `lm.3.10`, the analyst concluded the following: Since the model explains just more than 99.8% of the variability in Suggested Retail Price and the coefficient of Dealer Cost has a t-value greater than 412, the model `lm.3.10` is a highly effective model for producing prediction intervals for Suggested Retail Price.\[0.6em] % Provide a detailed critique of this conclusion.

Although the regression output in the summary of model `lm.3.10` *above looks good, there are several volations of the modeling assumptions revealed in following diagnostic plots:*



- *SuggestedRetailPrice, DealerCost and the residuals, are all skewed right*
- *The residuals have nonconstant variance*
- *The aggregation of the data around different lines in the plot suggests that some predictor variable(s) may be missing from the model*

Violations of the modeling assumptions undermine the validity of inferences we can make from the `summary(lm.3.10)` *output.*

Problem 1(b)

Carefully describe all the shortcomings evident in model `lm.3.10`. For each shortcoming, describe the steps needed to overcome the shortcoming.

Here are four possible shortcomings (you may have found others! Name any two legitimate criticisms, and ways to fix them, for full credit) *:*

- *The "SuggestedRetailPrice vs DealerCost", "Standardized Residuals vs DealerCost", and "Sqrt(|Standardized Residuals|) vs DealerCost" plots all show that Dealer Cost is substantially right-skewed. Right skewing tends to create high-leverage points in the data. A transformation of DealerCost to reduce the skewing would help: usually a fractional power, or a logarithm are good fixes for this.*
- *The Normal QQ Plot shows that the residuals are also right-skewed. The same sort of transformation (log or fractional power) of SuggestedRetailPrice will help to reduce this skewing.*

- *Both the Standardized Residuals plot and the "Sqrt(|Standardized Residuals|) vs DealerCost" plot show that the residuals have non-constant variance. You could suggest a variance-stabilizing transformation, or a log or power transformation, to help fix this problem.*
- *The Standardized Residual plot shows the data clustering along several different lines, suggesting that perhaps different vehicle types or brands have different relationships between SuggestedRetailPrice and DealerCost. One could explore this idea by considering an ANCOVA model, which we will talk about in later lectures.*

Problem 1(c)

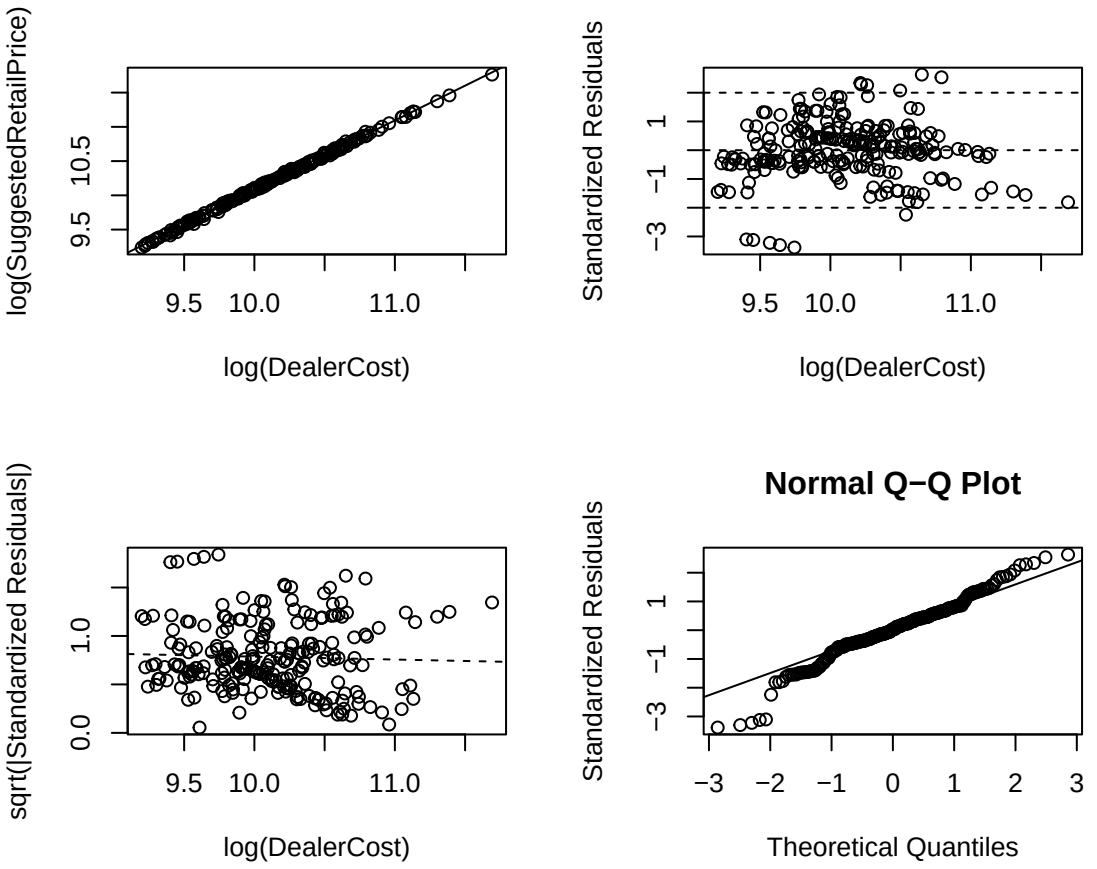
Is model `lm.3.11` below an improvement over model `lm.3.10` in terms of predicting Suggested Retail Price? If so, please describe all the ways in which it is an improvement.

```
lm.3.11 <- lm(log(SuggestedRetailPrice) ~ log(DealerCost),data=cars04)
summary(lm.3.11)
```

```
##
## Call:
## lm(formula = log(SuggestedRetailPrice) ~ log(DealerCost), data = cars04)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.062920 -0.008694  0.000624  0.010621  0.048798
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.069459   0.026459  -2.625  0.00924 **
## log(DealerCost)  1.014836   0.002616 387.942  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01865 on 232 degrees of freedom
## Multiple R-squared:  0.9985, Adjusted R-squared:  0.9985
## F-statistic: 1.505e+05 on 1 and 232 DF,  p-value: < 2.2e-16
```

`lm.3.11` *is definitely an improvement over* `lm.3.10`*. The regression output for both models is about the same, so the differences are in diagnostic plots below. Here are some ways in which the plots are better for* `lm.3.11` *than for* `lm.3.10` (You may have found other reasons. Name any two legitimate reasons for full credit).

- *Both DealerCost and SuggestedRetailPrice exhibit less right-skewing in* `lm.3.11` *than in* `3.10`*.*
- *While there are still outliers in the residual vs fitted plot for* `3.11`*, they are less extreme than for* `3.10`*.*
- *In addition to less right-skew, the normal QQ plot suggests more nearly-normal residuals for* `3.11` *than for* `3.10`*. We have pushed some values out into the left tail, but there are fewer of these in the QQ plot for* `3.11` *than there are for* `3.10`*.*



Problem 1(d)

Interpret the estimated coefficient of log(Dealer Cost) in `lm.3.11`.

As we know from the text, or from the handout "log xform and percent interpretation.pdf'' in the week03 folder in the files area, β_1 *is the expected percent change in* y *for a 1% change in* x*. Since* $\hat{\beta}_1 = 1.015$ *in* `summary(lm.3.11)`*, there is about a 1% change in SuggestedRetailPrice for every 1% change in DealerCost, according to the fitted model.*

Problem 1(e)

List any weaknesses apparent in `lm.3.11`.

Here are some weaknesses we can see in the diagnostic plots (you may have discovered others; list any two legitimate weaknesses to get full credit)*:*

- *From the QQ plot, both tails of the residual distribution are a bit long. Although the deviation is more impressive in the lower tail, there are more data points in the upper tail.*
- *The scale-location plot suggests that the variance of residuals may increase as DealerCost increases.*
- *Although it is not as evident in the diagnostic plots for* `lm.3.11` *as it is in the plots for* `lm.3.10`*, it still looks like the data aggregates along definable curves, which suggests that perhaps different car types have different relationships between DealerCost and SuggestedRetailPrice.*

(Aside: If we were to successfully model different DealerCost vs. SuggestedRetailPrice relationships for different car types, that might remove the other two problems above as well.)

Problem 2: Sheather, Ch 5, p. 147, #2.

We begin with a little exploration, and finding a good model.

```
houston <- read.csv("HoustonChronicle.csv",header=T)
str(houston)
```

```
## 'data.frame':    122 obs. of  5 variables:
##  $ District             : chr  "Alvin" "Alvin" "Angleton" "Angleton" ...
##  $ X.Repeating.1st.Grade: num  4.1 5.8 7.1 6.7 7.3 2.6 8.2 2.3 12.5 0 ...
##  $ X.Low.income.students: num  49.7 41.1 44.2 30.2 49.4 33.7 45.6 29.7 71.7 37.6 ...
##  $ Year                 : int  2004 1994 2004 1994 2004 1994 2004 1994 2004 1994 ...
##  $ County               : chr  "Brazoria" "Brazoria" "Brazoria" "Brazoria" ...
```

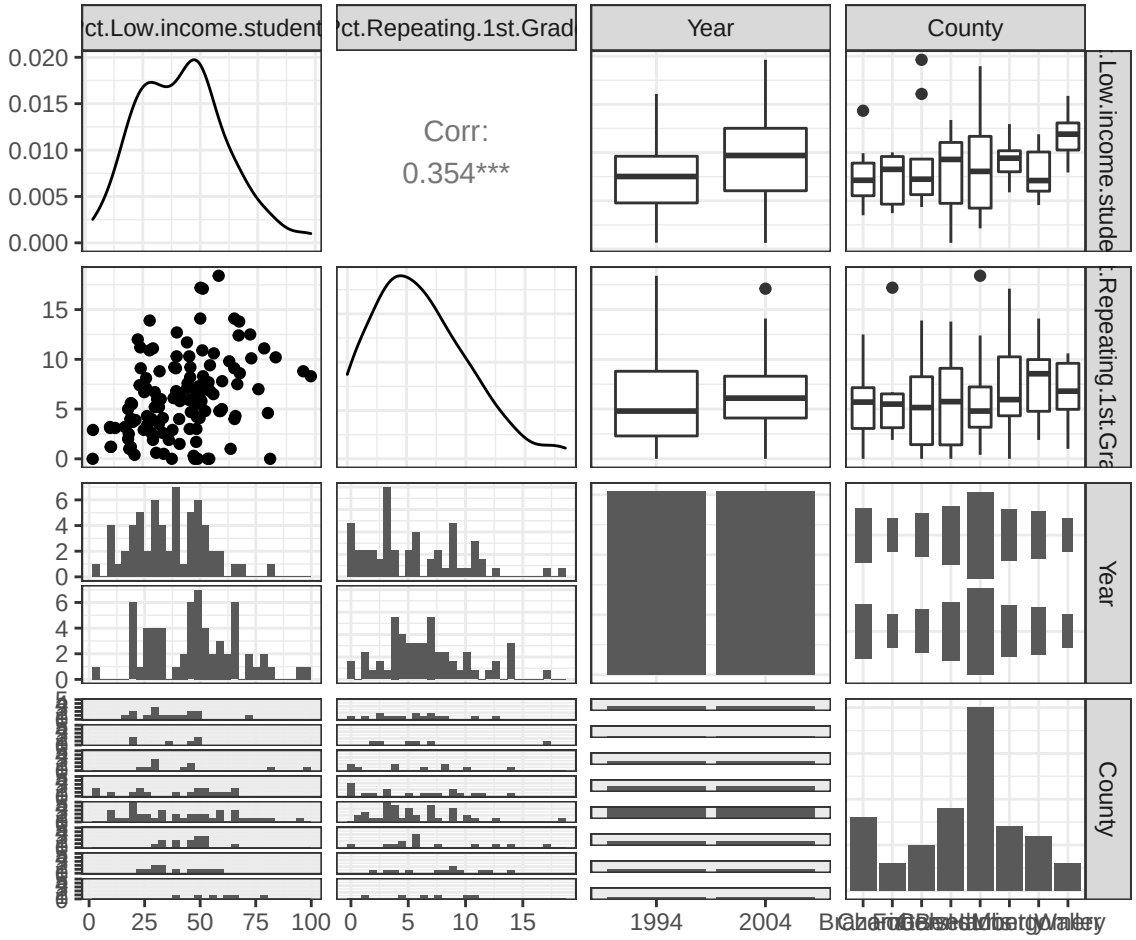
```
cat("\n")
```

```
for (i in names(houston)) {
  cat(i,"\n")
  print(summary(houston[,i]))
  cat("Number of unique values:",length(unique(houston[,i])),"\n\n")
}
```

```
## District
##    Length     Class      Mode
##       122 character character
## Number of unique values: 61
##
## X.Repeating.1st.Grade
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.100   5.700   6.076   8.750  18.400
## Number of unique values: 77
##
## X.Low.income.students
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.20   27.15   41.35   41.88   53.02   98.10
## Number of unique values: 111
##
## Year
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1994    1994    1999    1999    2004    2004
## Number of unique values: 2
##
## County
##    Length     Class      Mode
##       122 character character
## Number of unique values: 8
```

Apparently, there are only two "year" values: 1994 and 2004; so we will convert this to a factor variable, just to make interpretation of the coefficients easier; and we make a scatterplot matrix ("pairs" plot) of all of the variables except district. I'm also going to rename the two continuous variables to something more suggestive, and re-order the variables, so that the "pairs" plot makes a little more sense.

```
houston$Year <- as.factor(houston$Year)
names(houston)[c(2,3)] <- c("Pct.Repeating.1st.Grade", "Pct.Low.income.students")
houston <- houston[,c(1,3,2,4,5)]
ggpairs(houston[,-1])  # there are too many districts for the plot, so we omit it
```



There is some evidence in the "pairs" plots for all three of the hypotheses in parts (a), (b) and (c), but we now fit regression and ANCOVA models to see which effects are "significant" (i.e. large enough that they are not likely just due to noise in the data).

Problem 2(a) Low income associated with repeating first grade?

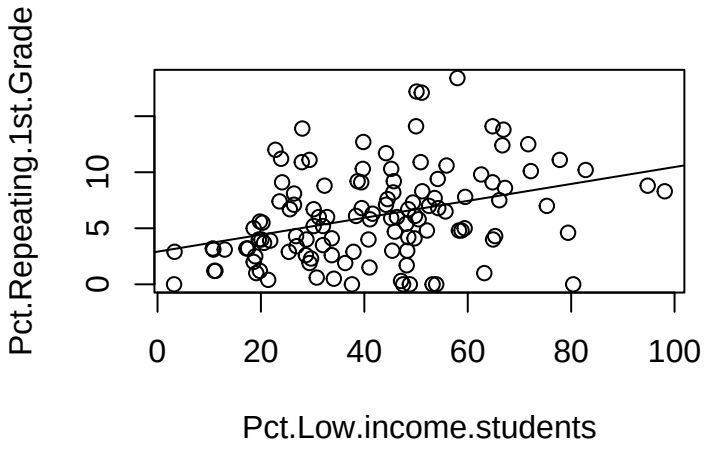
Let's start with a simple regression of repeating first grade on low income

```
summary(lm.2a <- lm(Pct.Repeating.1st.Grade ~ Pct.Low.income.students,data=houston))
```

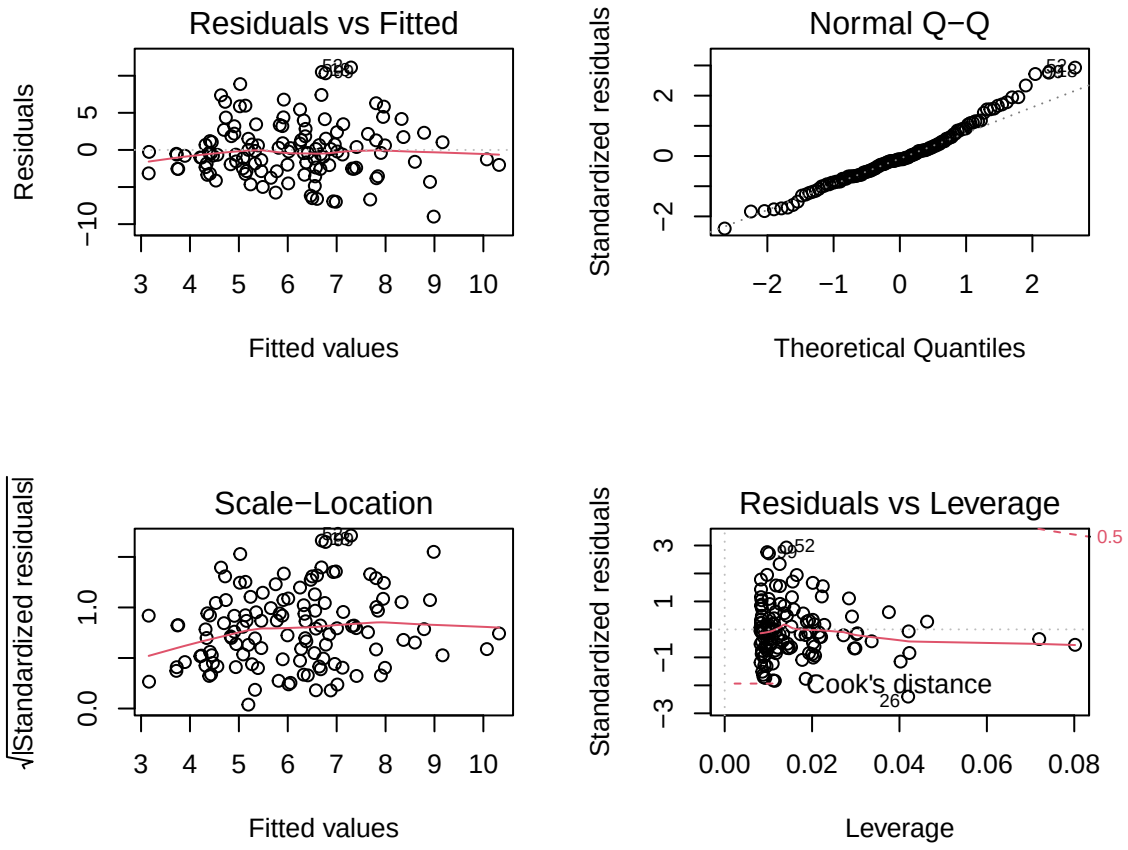
```
##
## Call:
## lm(formula = Pct.Repeating.1st.Grade ~ Pct.Low.income.students,
##     data = houston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9845 -2.5072 -0.4184  1.8505 11.1067
##
```

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.91419    0.83836   3.476 0.000709 ***
## Pct.Low.income.students   0.07550    0.01823   4.141 6.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.821 on 120 degrees of freedom
## Multiple R-squared:  0.125,  Adjusted R-squared:  0.1177
## F-statistic: 17.14 on 1 and 120 DF,  p-value: 6.472e-05
```

```
par(mfrow=c(1,1))
plot(Pct.Repeating.1st.Grade ~ Pct.Low.income.students,data=houston)
abline(lm.2a)
```




```
par(mfrow=c(2,2))
plot(lm.2a)
```



This actually looks like a good regression; the summary shows a very significant increasing association between percent of students repeating first grade and percent of students with low (family) incomes: the coefficient on percent low income students is $\hat{\beta}_1 = 0.07550$, with $SE(\hat{\beta}_1) = 0.01823$. The effect seems rather small, however—we expect an increase of only 0.08% of kids repeating first grade, for every 1% increase in kids in poverty.

Problem 2(b) More students repeating first grade in 2004-2005 than in 1994-1995?

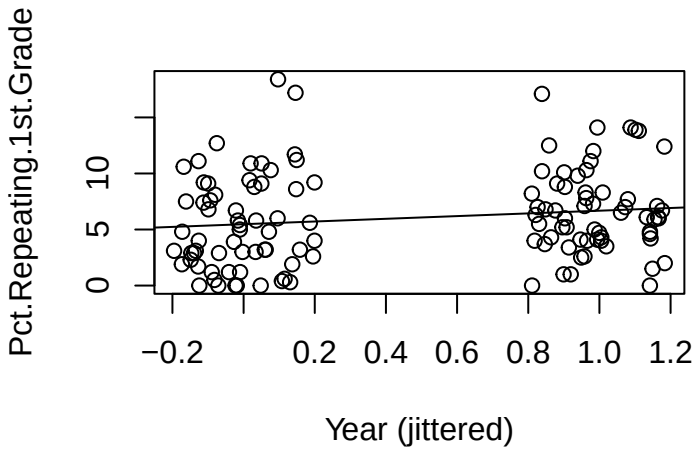
For this question, we will just regress repeating first grade on year

```
summary(lm.2b <- lm(Pct.Repeating.1st.Grade ~ Year,data=houston))
```

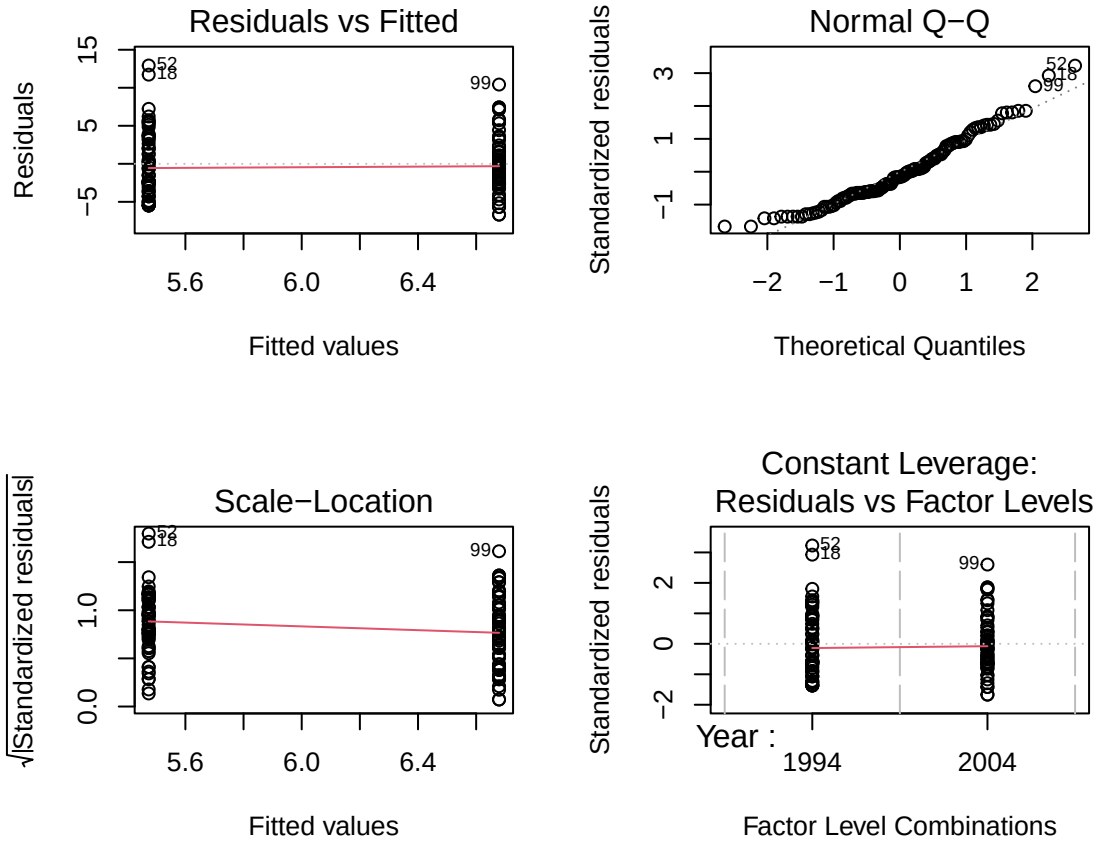
```
##
## Call:
## lm(formula = Pct.Repeating.1st.Grade ~ Year, data = houston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6787 -2.6537 -0.6262  2.5750 12.9262
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.4738     0.5172  10.584   <2e-16 ***
```

```
## Year2004      1.2049     0.7314   1.647    0.102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.039 on 120 degrees of freedom
## Multiple R-squared:  0.02212,    Adjusted R-squared:  0.01397
## F-statistic: 2.714 on 1 and 120 DF,  p-value: 0.1021
```

```
par(mfrow=c(1,1))
jy <- jitter(as.numeric(houston$Year)-1)
plot(Pct.Repeating.1st.Grade ~ jy,data=houston,xlab="Year (jittered)")
abline(lm.2b)
```




```
par(mfrow=c(2,2))
plot(lm.2b)
```



Again, this looks like a good regression model, in the sense that the assumptions underlying linear regression are approximately satisfied[1]. However, the estimated coefficient on Year, $\hat{\beta}_1 = 1.2$ with $SE(\hat{\beta}_1) = 0.73$ is not statistically significantly different from zero. From this we would conclude that there really is not enough evidence to say that there are more students repeating first grade in the 2004 school year than in 1994 school year.

Problem 2(c) Difference in the relationship of income with repeating between the two school years?

Now let's fit the interactive ANCOVA model. This will both help answer question (2c), and help us to understand whether our answers to questions (2a) and (2b) need any additional elaboration.

```
summary(lm.2c <- lm(Pct.Repeating.1st.Grade ~ Pct.Low.income.students*Year,data=houston))
```

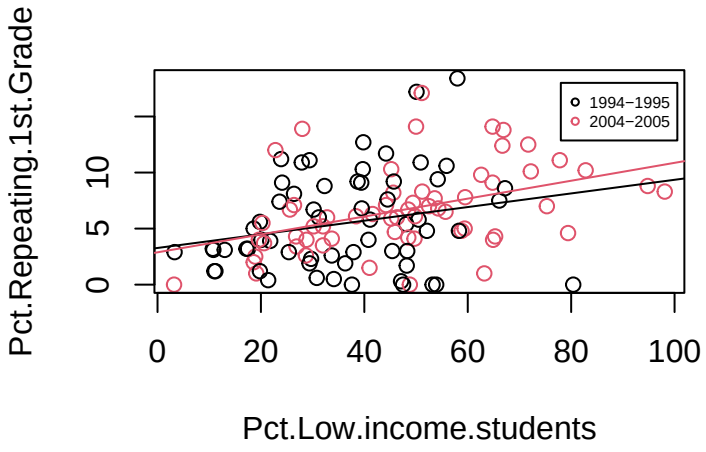
```
##
## Call:
## lm(formula = Pct.Repeating.1st.Grade ~ Pct.Low.income.students *
##     Year, data = houston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1606 -2.6121 -0.5576  1.7495 11.6014
```

[1]Remember, the clustering around different x values is just due to the fact that x is categorical. The important thing here is that we don't see any interesting patterns within or between those clusters.

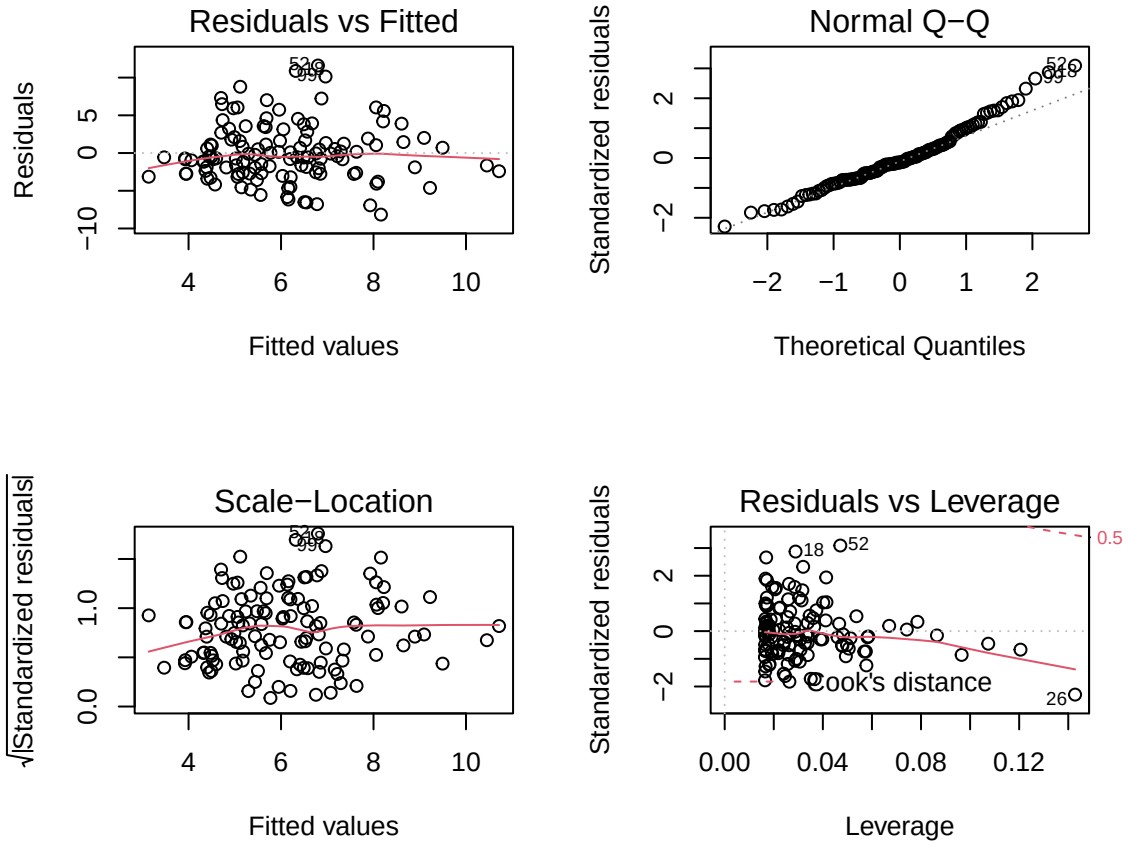
```
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       3.27194    1.22347   2.674  0.00855 **
## Pct.Low.income.students           0.06080    0.03093   1.966  0.05167 .
## Year2004                         -0.38956    1.76109  -0.221  0.82532
## Pct.Low.income.students:Year2004  0.01903    0.03949   0.482  0.63066
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.845 on 118 degrees of freedom
## Multiple R-squared:  0.1288, Adjusted R-squared:  0.1066
## F-statistic: 5.813 on 3 and 118 DF,  p-value: 0.0009689
```

The summary shows that neither the main effect for Year, nor the interaction between low income and Year, are significant in the model. This is illustrated in the scatterplot below: the regression lines for the two categories (year=1994-1995 and year=2004-2005) are nearly identical. The residual diagnostic plots that follow suggest that the modeling assumptions hold up fairly well here.

```
par(mfrow=c(1,1))
plot(Pct.Repeating.1st.Grade ~ Pct.Low.income.students,data=houston,
     col=as.numeric(houston$Year))
coefs <- coefficients(lm.2c)
abline(a=coefs[1],b=coefs[2],col=1)
abline(a=coefs[1]+coefs[3],b=coefs[2]+coefs[4],col=2)
legend(78,18,legend=c("1994-1995","2004-2005"),pch=1,col=c(1,2),cex=0.5)
```




```
par(mfrow=c(2,2))
plot(lm.2c)
```



We can also apply an F test using the ANOVA function in R to compare the full ANCOVA model in part (c) with the models in parts (a) and (b).

```
anova(lm.2a,lm.2c)
```

```
## Analysis of Variance Table
##
## Model 1: Pct.Repeating.1st.Grade ~ Pct.Low.income.students
## Model 2: Pct.Repeating.1st.Grade ~ Pct.Low.income.students * Year
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    120 1751.9
## 2    118 1744.4  2     7.512 0.2541 0.7761
```

Since the F statistic testing H_0: `Pct.Repeating.1st.Grade` $\sim$ `Pct.Low.income.students` vs H_A: `Pct.Repeating.1st.Grade` $\sim$ `Pct.Low.income.students` $*$ `Year` is nonsignificant ($F = 0.2541$, $p = 0.7761$), we cannot reject the simpler model in part (a): If we are starting with `Pct.Repeating.1st.Grade` $\sim$ `Pct.Low.income.students`, there is really no need to add `Year` (let alone an interaction with `Year`) to the model.

```
anova(lm.2b,lm.2c)
```

```
## Analysis of Variance Table
##
## Model 1: Pct.Repeating.1st.Grade ~ Year
## Model 2: Pct.Repeating.1st.Grade ~ Pct.Low.income.students * Year
```

```
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    120 1957.9
## 2    118 1744.4  2    213.52 7.2221 0.001099 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here the F statistic testing H_0: `Pct.Repeating.1st.Grade` $\sim$ `Year` vs H_A: `Pct.Repeating.1st.Grade` $\sim$ `Pct.Low.income.students` $*$ `Year` is highly significant ($F = 7.2221$, $p = 0.001099$), so we would reject the simpler model in part (b): starting with `Pct.Repeating.1st.Grade` $\sim$ `Year`, we really do get a better model by adding `Pct.Low.income.students` to the model.

Note that we cannot perform an F test like `anova(lm.2a, lm.2b)` directly[2], since these models are not nested. We can infer, informally, from the above tests though, that of all three models, `Pct.Repeating.1st.Grade` $\sim$ `Pct.Low.income.students` seems to be the best. Thus, there is no difference between years in the way that low income is related to repeating first grade.

Problem 3: Sheather, Ch 6, pp. 224 ff., #5.

- For part (b), feel free to transform variables as needed so that (i) the assumptions of the linear regression model are better satisfied; and (ii) the model is still "explainable" to non-statisticians.
- For part (e): We will discuss variable selection formally next week. For now, remember that the t statistic for each column X_j of the X matrix tests whether the coefficient β_j is significantly different from zero, *after including all other columns of X in the model.*

Problem 3(a) Transform PrizeMoney $\rightarrow$ log(PrizeMoney) but no other transformations?

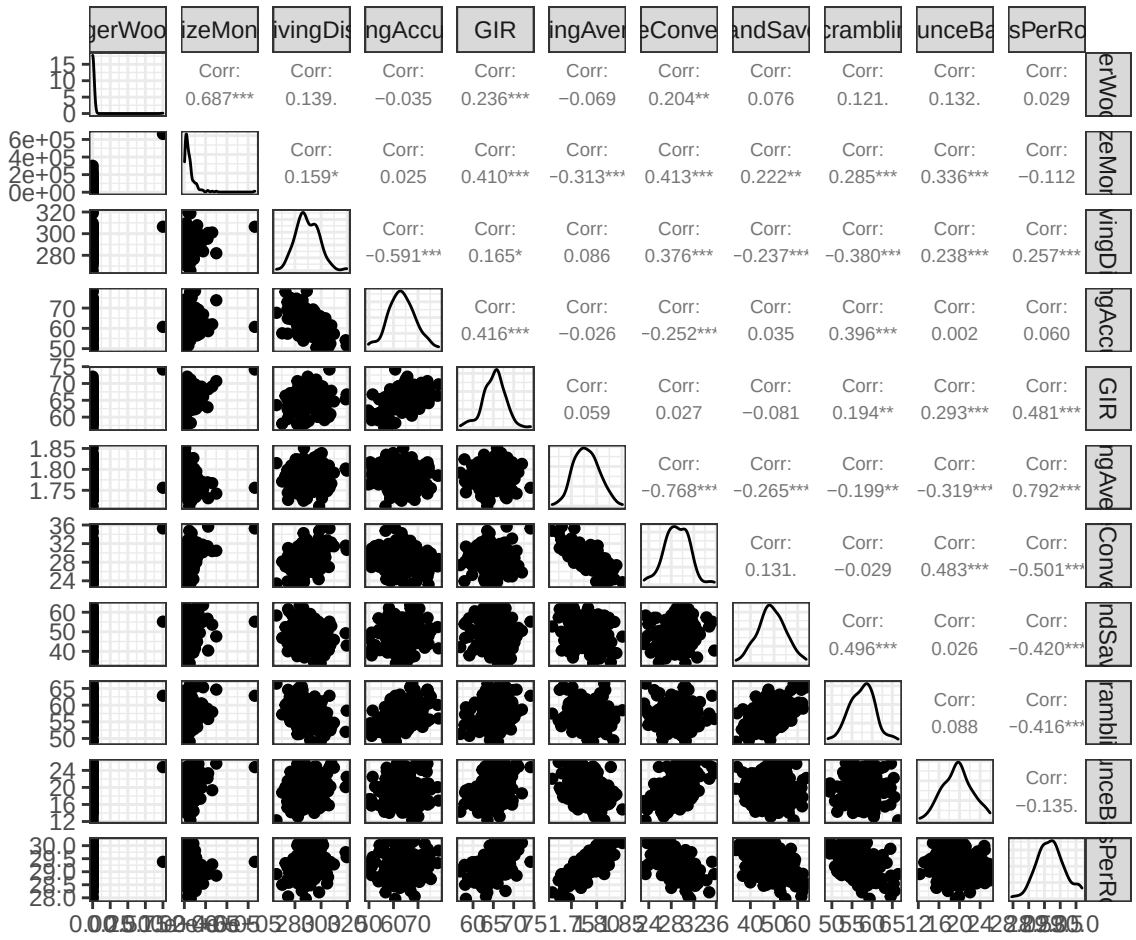
From the `ggpairs` *plot below, and possibly from futher analysis that you did for Quiz 2, only the variables* `TigerWoods` *and* `PrizeMoney` *are strongly skewed. The variable* `TigerWoods` *is a dummy/indicator variable, so we won't transform that. Because of the strong skew in* `PrizeMoney`*, a log transformation makes sense for this variable (Box-Cox would confirm that, by suggesting a power of* `PrizeMoney` *very close to zero).*

```
golf <- read.csv("pgatour2006.csv",header=TRUE)
names(golf)
```

```
##  [1] "Name"               "TigerWoods"         "PrizeMoney"
##  [4] "AveDrivingDistance" "DrivingAccuracy"    "GIR"
##  [7] "PuttingAverage"     "BirdieConversion"   "SandSaves"
## [10] "Scrambling"         "BounceBack"         "PuttsPerRound"
```

```
ggpairs(golf[-1],upper = list(continuous = wrap("cor", size=2.5))) ## don't include golfer's name
```

[2]There are ways to compare non-nested models, e.g. cross-validation, AIC, BIC, etc., but it doesn't work with F tests or likelihood ratio tests.



None of the other variables show strong skew visually, though Box-Cox might recommend various transformations for them. In the interest of being able to communicate clearly about any predictors in the model, and since these predictors don't exhibit much skew anyway, I'm going to leave them untransformed.

Problem 3(b) Develop a valid full regression model containing all seven potential predictors (listed in the problem stmt), using log(PrizeMoney) as the response variable

If we look at the data set directly

```
oldwidth <- options()$width
options(width=200)
head(golf)
```

```
##              Name TigerWoods PrizeMoney AveDrivingDistance DrivingAccuracy   GIR PuttingAverage BirdieConversion SandSaves Scrambling BounceBack PuttsPerRound
## 1   Aaron Baddeley          0      60661              288.3           60.73 58.26          1.745            31.36     54.80      59.37      19.30          27.96
## 2       Adam Scott          0     262045              301.1           62.00 69.12          1.767            30.39     53.61      57.94      19.35          29.28
## 3     Alex Aragon           0       3635              302.6           51.12 59.11          1.787            29.89     37.93      50.78      16.80          29.20
## 4       Alex Cejka          0      17516              288.8           66.40 67.70          1.777            29.33     45.13      54.82      17.05          29.46
## 5      Arjun Atwal          0      16683              287.7           63.24 64.04          1.761            29.32     52.44      57.07      18.21          28.93
## 6 Arron Oberholser          0     107294              285.0           62.53 69.27          1.775            29.20     47.20      57.67      20.00          29.56
```

```
options(width=oldwidth)
```

we see that each line is for one golfer, and each golfer has a different name. Thus, if we include `Name` *in the regression (a) the model will fit perfectly, but (b) we will get no information about the other predictors (try it! it's truly ugly!).*

So I will remove `Name` *from the data set and proceed from there...*

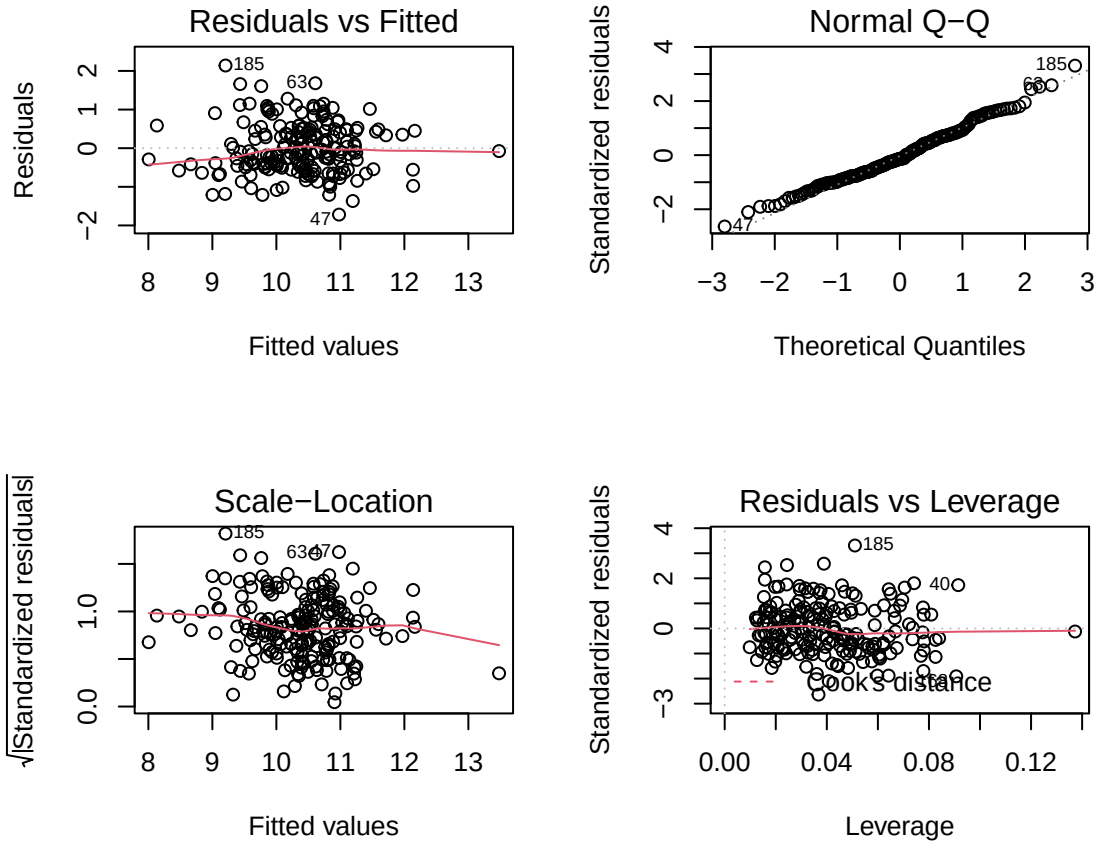
```
golf <- golf[,-1]
```

Really, nothing more than this model is required:

```
summary(lm.3.1 <- lm(log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage
                     + BirdieConversion + SandSaves + Scrambling + PuttsPerRound,
                     data=golf))
```

```
##
## Call:
## lm(formula = log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage +
##     BirdieConversion + SandSaves + Scrambling + PuttsPerRound,
##     data = golf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71949 -0.48608 -0.09172  0.44561  2.14013
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.194300   7.777129   0.025 0.980095
## DrivingAccuracy  -0.003530   0.011773  -0.300 0.764636
## GIR               0.199311   0.043817   4.549 9.66e-06 ***
## PuttingAverage   -0.466304   6.905698  -0.068 0.946236
## BirdieConversion  0.157341   0.040378   3.897 0.000136 ***
## SandSaves         0.015174   0.009862   1.539 0.125551
## Scrambling        0.051514   0.031788   1.621 0.106788
## PuttsPerRound    -0.343131   0.473549  -0.725 0.469601
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6639 on 188 degrees of freedom
## Multiple R-squared:  0.5577, Adjusted R-squared:  0.5412
## F-statistic: 33.87 on 7 and 188 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(lm.3.1)
```



You can play with other transformations, but this is already a pretty good model, in terms of the casewise diagnostic plots, and so I wouldn't bother (this is easier to explain to a client or collaborator!):

- *The residuals vs fitted values plot does not show any pattern other than smaller amounts of data for very low or very high $\hat{y}$'s.*
- *The normal QQ plot looks very good except for a slight hint of a bit of a long tail to the right, but this is very minor and probably can be ignored.*
- *The Scale-Location plot is almost a perfect example of constant variance, with just a little bit of small-sample bias for larger $\hat{y}$ where there is very little data*
- *The residuals vs leverage plot shows no concerning points in the northeast or southeast corners of the plot.*

Given that the model `lm.3.1` *seems to satisfy the assumptions of linear regression well, we can use the overall F test to see that regressing on these variables is better than the intercept-only model ($F = 33.872$ on 7 and 188 df, $p < 2.2 \times 10^{-16}$). Here R^2 is a fairly modest 0.66, and only two of the seven predictors have significant t statistics.*

Problem 3(c) Identify any points that should be investigated. Give one or more reasons to support each point chosen.

One point in the casewise diagnostic plots seems to have a high fitted value and high leverage. One way to find that point is this (sorry for the small type; trying to make a row of the data frame fit on one line)...

```
oldwidth <- options()$width
options(width=200)

max.fitted.val <- max(fitted(lm.3.1))
golf[fitted(lm.3.1)==max.fitted.val,]
```

```
##     TigerWoods PrizeMoney AveDrivingDistance DrivingAccuracy   GIR PuttingAverage BirdieConversion SandSaves Scrambling BounceBack PuttsPerRound
## 178          1     662771              306.4           60.71 74.15          1.756            35.26     55.17      62.81      24.77         29.38
```

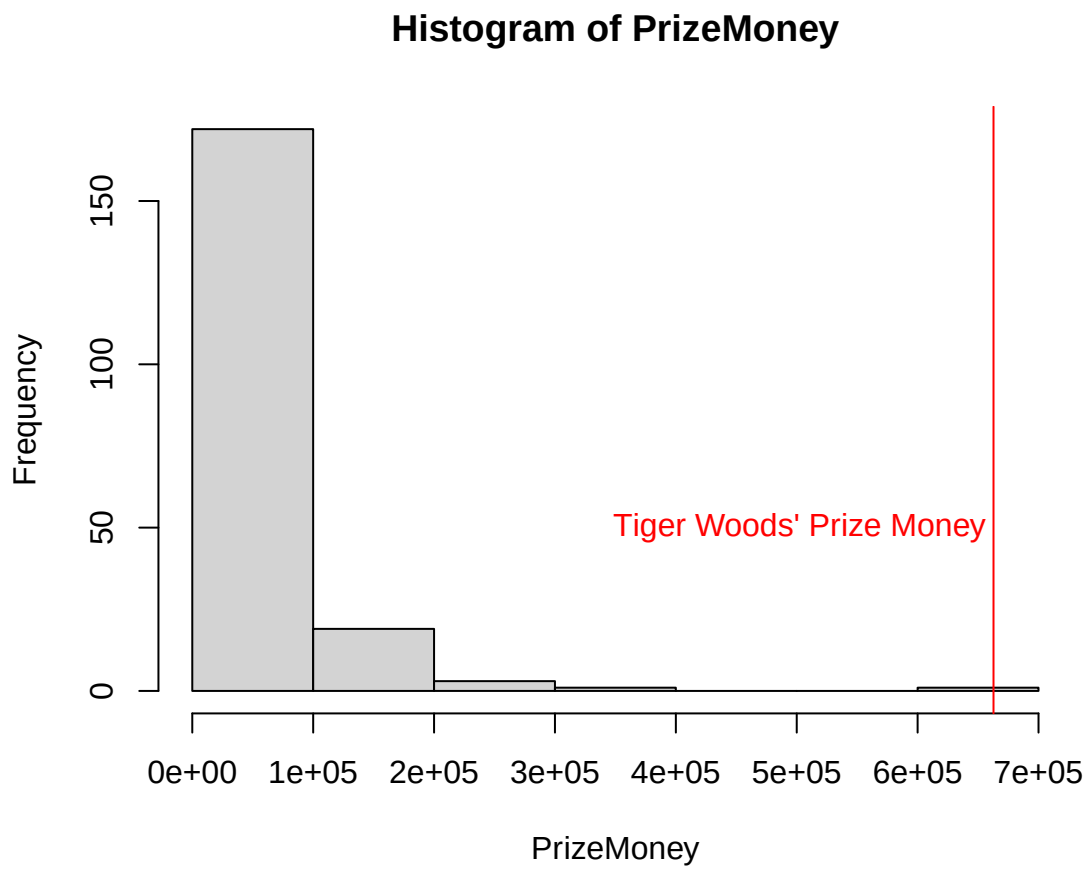
```
max.leverage <- max(hatvalues(lm.3.1))
golf[hatvalues(lm.3.1)==max.leverage,]
```

```
##     TigerWoods PrizeMoney AveDrivingDistance DrivingAccuracy   GIR PuttingAverage BirdieConversion SandSaves Scrambling BounceBack PuttsPerRound
## 178          1     662771              306.4           60.71 74.15          1.756            35.26     55.17      62.81      24.77         29.38
```

```
options(width=oldwidth)
```

. . . and we find that the unusual data point belongs to Tiger Woods! We already know there is a special indicator variable for Tiger Woods, which is perhaps not surprising since he was far and away the top prize money earner in golf in 2006:

```
attach(golf)
tw <- PrizeMoney[178]
hist(PrizeMoney)
abline(v=tw,col="red")
text(tw-160000,50,"Tiger Woods' Prize Money",col="red")
```




```
detach(golf)
```

Problem 3(d) Describe any weaknesses in your model.

As shown above, the model seems to satisfy the assumptions of linear regression, so the overall fit seems good. Weaknesses of the model include:

- $R^2 = 0.66$ *so only 66% of the variation in* `log(PrizeMoney)` *is explained by the predictors.*
- *Only two of the seven predictors (*`GIR` *and* `BirdieConversion`*) have coefficients significantly different from zero, according to the t statistics.*

Problem 3(e) The golf fan wants to remove all predictors with insignificant t -values from the full model in a single step. Explain why you would not recommend this approach.

The t-test for a predictor is the numerical equivalent of the added-variable plot: it shows the importance of that predictor, after other predictors have been added to the model. Indeed, $t^2 = F$, the partial F statistic for adding that predictor, after all other predictors have been added to the model.

If two or more predictors are co-linear, the t statistics for any of them—even for all of them—can be nonsignificant, even though they contribute significantly to explaining y.

Thus, removing all the predictors with non-significant t-statistics at once can remove important variables in the model. As seen in the `ggpairs` *plot in part (a), some predictors are highly correlated with one another, and linear combinations of them may be even more highly correlated with one another.*

If we are going to use t statistics to eliminate variables, it is better to eliminate one-at-a-time, starting with the t statistics closest to zero (unless our client or collaborator wants to eliminate something else first). Each time we eliminate a variable, we would re-fit the model and discover that all the remaining t-statistics have changed, and we would use the new t-statistics to select the next predictor to eliminate.

This recipe is easy to follow but it does not necessarily lead to a great model, because (a) it is "greedy" in that it only looks at the "worst" variable at each step, rather than considering other orders in which to eliminate variables, and (b) it re-uses the data many times, and so is vulnerability to "capitalization on chance": t statistics in subsequent models after removing one or more predictors do not have the "textbook" distributions, and we may be making severe errors by relying on them for variable selection.

In Chapter 7 and related readings, we will look at better ways to do variable selection.

Problem 4

In the folder for this hw assignment you will find a pdf called "COVID breakthrough rates in England". This is a recent article from the medical journal *The Lancet*. Note that the article is exactly in IMRAD format, with sections labelled "Introduction", "Methods", "Results" and {"Discussion".

The Abstract for this paper has a special form, called a "structured abstract", in which there is a short, labelled, paragraph corresponding to each section of the paper. We will not be writing structured abstracts in our class, but it is good to see the additional information that a structured abstract can contain. In our class, we will be writing shorter abstracts, consisting of approximately one sentence per section of the paper (so, 4–5 sentences for a paper with 4 sections). Each sentence highlights the main point of each section, possibly with one extra sentence giving the main resul of the paper.

Problem 4(a)

Write a one-paragraph abstract with exactly four sentences, one for each section of the paper: Introduction, Methods, Results and Discussion. Each sentence should highlight the main point of each section, and together the four sentences should tell the story of the paper. The last sentence should include the main result of the paper (or, if you need a fifth sentence to give the main result, that is fine too).

(For most of this abstract I simply copied or merged sentences from the structured abstract on p. 1 of the article [you could of course write your own sentences summarizing each part of the article]. I thought it would be helpful if I added a footnote for the place I got each sentence [you do not have to do this for your answer].)

This study aimed to identify risk factors for post-vaccination SARS-CoV-2 infection and describe the characteristics of post-vaccination illness[3]*. We used univariate logistic regression models (adjusted for age, BMI, and sex) to analyse the associations between risk factors and post-vaccination infection, and the associations of individual symptoms, overall disease duration, and disease severity with vaccination status, in self-report data from UK-based, adult (≥ 18 years) users of the COVID Symptom Study mobile phone app*[4]*. Vaccination (compared with no vaccination) was associated with reduced odds of hospitalisation or having more than five symptoms in the first week of illness following the first or second dose, and long-duration (≥ 28 days) symptoms following the second dose; following first dose only, older adults, individuals living in deprived areas, and obese individuals all experienced higher risk of breakthrough infection*[5]*. Our findings might support caution around relaxing physical distancing and other personal protective measures in the post-vaccination era, particularly around frail older adults and individuals living in more deprived areas, even if these individuals are vaccinated, and might have implications for strategies such as booster vaccinations*[6]*.*

Problem 4(b)

Does the paper appropriately address each of the parts of an IMRAD paper as described the "IMRAD: What goes into each section" pdf? (in the hw01 folder in the files area of our Canvas site.)

For each section below, either say "yes this section has the right content", or say "no" and describe what is missing and/or what needs to be moved to another section of the paper or deleted.

Note: *I give one possible set of answers below. You do not have to give the same answers, as long as the answers you give are thoughtful.*

- *Your "yes" and "no" answers do not have to match mine.*
- *If you say "yes the content in this section is fine" you do not have to give examples (even though I do below)*
- *If you say "no" then we want to see thoughtful examples/discussion of what is missing or what needs to be moved elsewhere, etc.*

Introduction: *Yes this section has the right content* (you do not need to give examples; I did just so you could see where some of the elements are)

- *Supplies sufficient background information, e.g.:*

  > *"Vaccination against SARS-CoV-2 is a leading strategy to change the course of the COVID-19 pandemic worldwide. . . "*

  > *"A previous analysis of community-based individuals in the COVID Symptom Study showed a significant reduction in infection post-vaccination from 12 days after the first dose. . . "*

  > *"Nonetheless, some people still contract COVID-19 after vaccination, and further virus variants could evolve. . . "*

- *Shows Define lacunae and shortcomings in current state of knowledge & Rationale for the study, e.g.:*

  > *Individuals with COVID-19 have differing symptoms and clinical needs.20 Elucidating symptom profiles in individuals with COVID-19 after vaccination has clinical utility,*

[3]Last sentence of summary paragraph, p. 1
[4]Combining first and last sentences of Methods paragraph, p. 1
[5]Second-to-last sentence of Findings paragraph, combined with a summary of the risk factor analysis.
[6]Last sentence of Interpretation paragraph, p. 1

facilitating the identification of risk groups for intervention, predicting medical resource requirements, and informing appropriate testing guidelines. Additionally, some unvaccinated individuals with COVID-19 have prolonged illness duration (so-called long COVID) and whether vaccination reduces the risk of long COVID is currently unknown.

- *States aim of the study, e.g.:*

  Therefore, we aimed to (1) describe individual risk factors associated with SARS-CoV-2 infection at least 14 days after first vaccination or 7 days after second vaccination, and (2) assess illness duration, severity, and symptom profile in individuals with SARS-CoV-2 infection after their first and second vaccinations, compared with unvaccinated individuals with SARS-CoV-2 infection.

Methods: *Yes this section has the right content* (again, you do not need to give examples; I did just so you could see where some of the elements are)

- *Broken down into subsection (may not be necessary for all papers, but nice here):*
  - *Study design and participants*
  - *Risk factor variable definitions*
  - *Disease severity, duration, and symptom definitions*
  - *Statistical analysis*
  - *Role of the funding source*

  These sections provide all of the elements recommended by the slides on what goes into an IMRAD paper (study design, variable definitions, how outcomes are measured, analysis and statistical methods).

  Notes:

  1. *In an IDMRAD (Introduction, Data, Methods, Results, and Discussion) paper (which we will be writing later in the course, instead of IMRAD papers), the first three bullets above (study design, variable defintions) would be moved into the Data section of the paper, between the Introduction and Methods sections of the paper.*
  2. *The subsection "Role of the funding source" would not be needed in our papers, but it is needed in papers where the author needs to show that their research was not unduly influenced by who funded the work. This is actually an aspect of addressing **ethics** in the study. (Other aspects of ethics include: how fairly and respectfully did the researchers treat human subjecs; whether invidually-identifiable results were kept confidential, whether related research by other researchers was fairly credited and cited in the paper, etc.)*
  3. *If the paper is intended to answer more than one research question, another way to organize the Methods section would be to have one subsection for each research question. The Results section would then have the same subsections. Since this study had two major questions — (1) describe individual risk factors associated with SARS-CoV-2 infection at least 14 days after first vaccination or 7 days after second vaccination; and (2) assess illness duration, severity, and symptom profile in individuals with SARS-CoV-2 infection after their first and second vaccinations, compared with unvaccinated individuals with SARS-CoV-2 infection — another way to organize this paper would have been to have two subsections, one for each question, in the Methods section, and in the Results section. Organizing these sections by research question makes it more skimmable by readers who are in a hurry (and most readers are!).*

Results: *Yes this section has the right content, **BUT** I thought this was the weakest section of the paper* (again, you do not need to give examples; I did just so you could see where some of the elements are)

- *The first part of the discussion talks about the results of the experiment/data collection, which was good.*
- *The next several paragraphs discuss statistical analyses in a natural order, but they are not very good at helping the reader remember which of the two major questions each analysis goes with*
- *Here, subsections labelled with each research question would be really helpful.*
- *Tables 1 and 2 and Figures 1,2,3,4 are very helpful in answering the first research question (if the reader can remember what it was!) but there doesn't seem to be an easily identifiable summary of results for the second research question (results for the second question are discussed at the end of p. 7 and the first column of page 9, in text form only).*

A good rule is: if the reader has to work too hard to find what they are looking for in the paper, then the paper really should be better written or organized!

Discussion: *Yes this section has the right content* (you do not need to give examples; I did just so you could see where some of the elements are)

- *A summary (recapitulation) of the study and its major findings appears on pp 8–10.*
- *Limitations and strengths of the study are discussed on p. 11*
- *Broader implications are discussed in the last paragraph of the paper.*
- *It might have been nice for the authors to indicate ideas for future research building on this work, as well.*