Homework 05 Solutions

10/4/2021

1. (Sheather 6.7.5)

(a) Agree. Since the Y variable is heavily right-skewed and ranges over several orders of magnitude, a log transformation probably makes sense. To confirm this, you can regress `PrizeMoney` against the other predictors and then regress `log(PrizeMoney)` against the other predictors. The residuals are clearly non-normal in the first case and are approximately normal in the second case, which suggests that the log transformation of the outcome variable is appropriate. There is no apparent need to transform the predictors.

(b) I simply regressed the log-transformed `PrizeMoney` variable (`LogPrizeMoney`) against the seven predictors. The diagnostic plots (Figure 2) look about as good as one could ask. Since I know nothing about golf, I have no reason to expect any particular interactions, so for the sake of simplicity, and since the diagnostics look good, I will leave the model as is. A summary of the model is below, and Figure 1 contains a pairs plot of the data. Some collinearity is evident in the pairs plots, in particular between `PuttingAverage` and `BirdieConversion` and between `PuttingAverage` and `PuttsPerRound`.

It is tempting to remove some of the predictors with high p-values, but until we are working with a golf expert as a collaborator we do not know which ones to remove. If we wanted to remove some predictors without further information, we might consider removing ones for which the 95% confidence interval for $\hat{\beta}$ (a) is narrow and (b) contains zero. `DrivingAccuracy`, `SandSaves` and maybe `Scrambling` could be considered for removal. On the other hand `PuttingAverage` has a very wide confidence interval, which suggests that we may not have enough information yet to have a good estimate of its $\hat{\beta}$. (See also the comments after part (e) below.)

Summary of the golf model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1943	7.777	0.02498	0.9801
DrivingAccuracy	-0.00353	0.01177	-0.2998	0.7646
GIR	0.1993	0.04382	4.549	9.658e-06
PuttingAverage	-0.4663	6.906	-0.06752	0.9462
BirdieConversion	0.1573	0.04038	3.897	0.0001355
SandSaves	0.01517	0.009862	1.539	0.1256
Scrambling	0.05151	0.03179	1.621	0.1068
PuttsPerRound	-0.3431	0.4735	-0.7246	0.4696

Observations	Residual Std. Error	R^2	Adjusted R^2
196	0.6639	0.5577	0.5412

(c) There are no points that clearly require investigation. Point 185 has a large residual, but in a data set of size 196 it is not particularly surprising to see a point with a standardized residual of 3 or so. There are no bad leverage points, and since the diagnostics look good in general, there's nothing obvious to investigate.

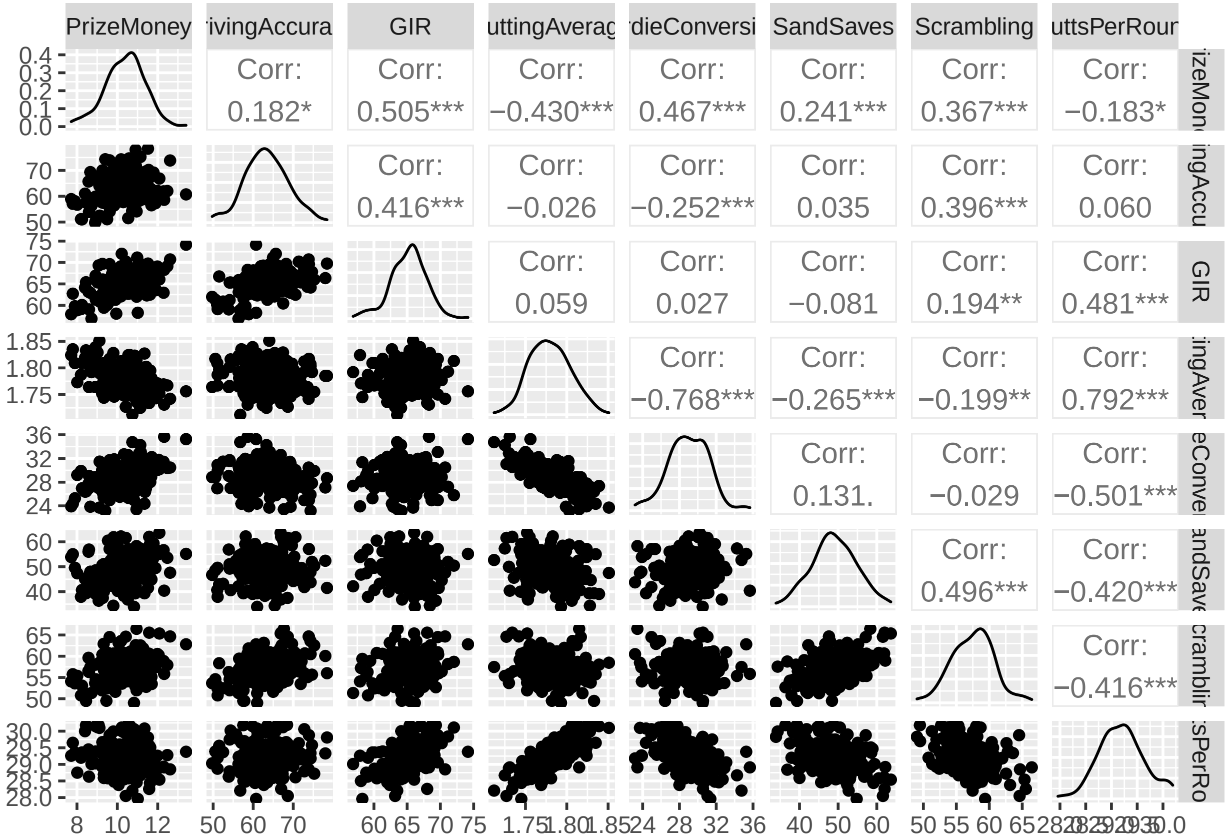

Figure 1: Pairs plot of golf data, with log transformation of PrizeMoney variable.

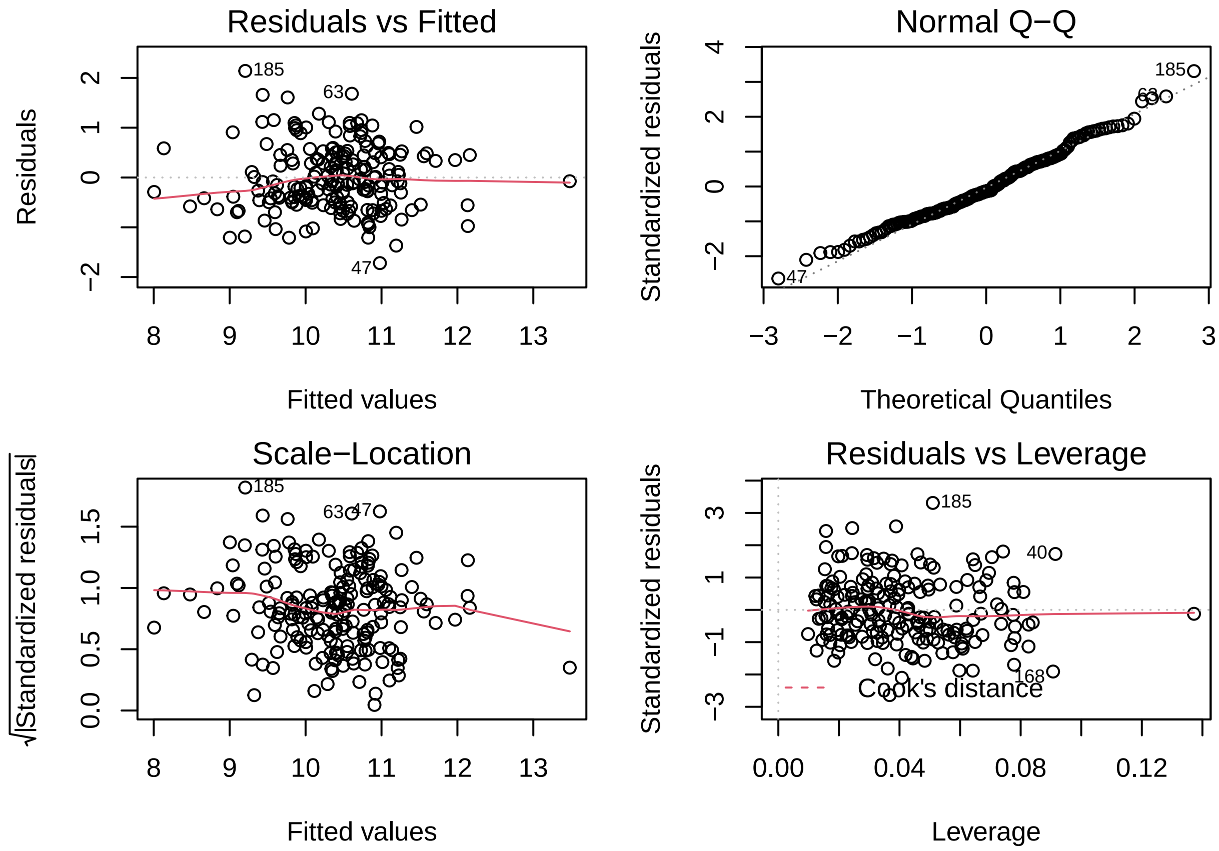

Figure 2: Problem 1: Diagnostic plots.

(d) As noted above, there's some collinearity in the predictors, so some of the predictors may be redundant; and the standard errors of the estimates may be large because of this. As always, the assumption that the outcome is linearly related to the predictors is a strong one, and it means the model will not capture nonlinearities.

(e) There are two problems with relying only on insignificant t-values to eliminate predictors in a model. The first is that a t-value can be small because the predictor is (nearly) collinear with another predictor or group of predictors. In that case, it's not obvious which predictor (if any!) to eliminate, without a conversation with a subject matter expert. The second is that hypothesis tests are asymmetrical, in the sense that when we reject the test, we are making an assertion that H_0 is false (with the attendant possibility of Type I error), whereas when we fail to reject the test, it could be because we have strong evidence that H_0 is true or because we simply don't have enough evidence to make a judgment either way.

> A TA from a previous iteration of the class comments further on the second reason:
>
> The value of the t-statistic associated with a coefficient β_k in a regression is a function of the magnitude of β_k, the variance of the predictor X_k, the amount of noise around the regression line, and the sample size. We may be able to reject the null for coefficients with very small magnitudes if, say, the variance of the associated predictor is high; conversely, we may fail to reject the null for coefficients with large magnitudes because the range of the associated predictor is small. Just because a test is insignificant does not mean that the predictor is insignificant in a real world sense. (If, however, the confidence interval for the predictor were very small and contained 0, then we might be able to confidently say that the coefficient was insignificant in a real world sense, as long as the model were generally credible.) For more details, see Professor Shalizi's lecture notes from 36-401:
>
> - Fall 2015, Lecture 8. Statistical Signifcance: Uses and Abuses (esp. Sections 5, 6, & 7)
> - Fall 2015, Lecture 15. What, Exactly, Is R Testing?
> - Why Variable Selection Using p-Values Is a Bad Idea

2. The Beauty Data Again

(a)

Why might we want to transform a predictor? There are at least three reasons:

- For interpretability, for example if we want the intercept to have a particular meaning, or if we want to be able to think about percentage changes in y relative to x, rather than linear changes.
- To reduce the leverage of particular points.
- To satisfy modeling assumptions, for example if y is not linearly related to x but is linearly related to some transformation $g(x)$.

Now let's think about the predictors we have in this problem. The distributions of all the non-indicator predictors are given in Figure 3.

Regarding the first item 1, I see no compelling reason to transform the predictors for the sake of interpretability. True, a y-value with respect to an "age" of 0, for example, will not be meaningful, but we won't ever be predicting in this range anyway.

With respect to the second item above, there are some predictors with very skewed distributions, so we could consider transforming them to reduce the skewness. In particular, `btystdvariance`, `didevaluation`, and `students` are strongly right-skewed and strictly positive, so we could consider log-transforming them.

We also have `btystdavepos`, which is strongly right-skewed, and `btystdaveneg`, which is strongly left-skewed. However, both these variables have a large number of values which are 0, so obviously no deterministic transformation will change the fact that there are a large number of instances with the same value. Since the rest of the values are fairly spread out within their ranges, transformations won't necessarily be useful.

With respect to the third item above, we'd need to fit the model and then check whether the assumptions appear reasonable.

(b)

I tried fitting the model with no transformations as well as with log transformations of `btystdvariance`, `didevaluation`, and `students`. The diagnostics for the former (Figure 4) looked slightly better than the latter, so I will stick with the model with no transformations. The t-values and VIF values are in Table 3.

```
##    Length     Class      Mode
##         1 knit_asis character
```

Table 3: Problem 2b: Table of t-values and Variance Inflation Factors for the model coefficients.

	t-value	VIF
(Intercept)	0.028	0.000000e+00
age	2.352	2.170000e+00
beautyf2upper	0.996	3.809736e+13
beautyflowerdiv	-0.673	6.305091e+13
beautyfupperdiv	1.017	1.036440e+03
beautym2upper	-1.067	1.239121e+14
beautymlowerdiv	0.207	7.168414e+13
beautymupperdiv	1.853	5.091113e+13
blkandwhite	0.222	1.899000e+00
btystdave	-0.820	6.244059e+13
btystdaveneg	-1.177	1.451961e+11
btystdavepos	-1.177	1.856450e+11
btystdf2u	-0.725	4.145703e+13
btystdfl	0.808	7.379827e+13
btystdfu	0.882	2.734726e+12
btystdm2u	1.167	1.300686e+14
btystdml	-0.032	6.617188e+13
btystdmu	-1.655	5.006012e+13
btystdvariance	-0.642	1.699000e+00
didevaluation	-1.201	4.525400e+01
female	-1.205	2.286000e+00
formal	0.896	1.541000e+00
fulldept	-0.197	1.772000e+00
lower	0.325	1.701000e+00
minority	-1.318	1.578000e+00
nonenglish	-1.674	1.497000e+00
onecredit	1.827	1.751000e+00
percentevaluating	2.443	2.685000e+00
profevaluation	48.783	1.379000e+00
students	1.105	5.096800e+01
tenured	0.449	2.720000e+00
tenuretrack	-1.019	2.605000e+00

(c) Using the rule of thumb that a VIF above 5 is high, many of the predictors have extremely high VIFs. If the purpose of the model is to perform inference on the coefficients, then you may wish to eliminate some of these predictors. If the purpose of the model is to generate accurate predictions, then you may wish to leave them in, though we would need to evaluate predictive performance on various subsets of predictors in order to make that decision.

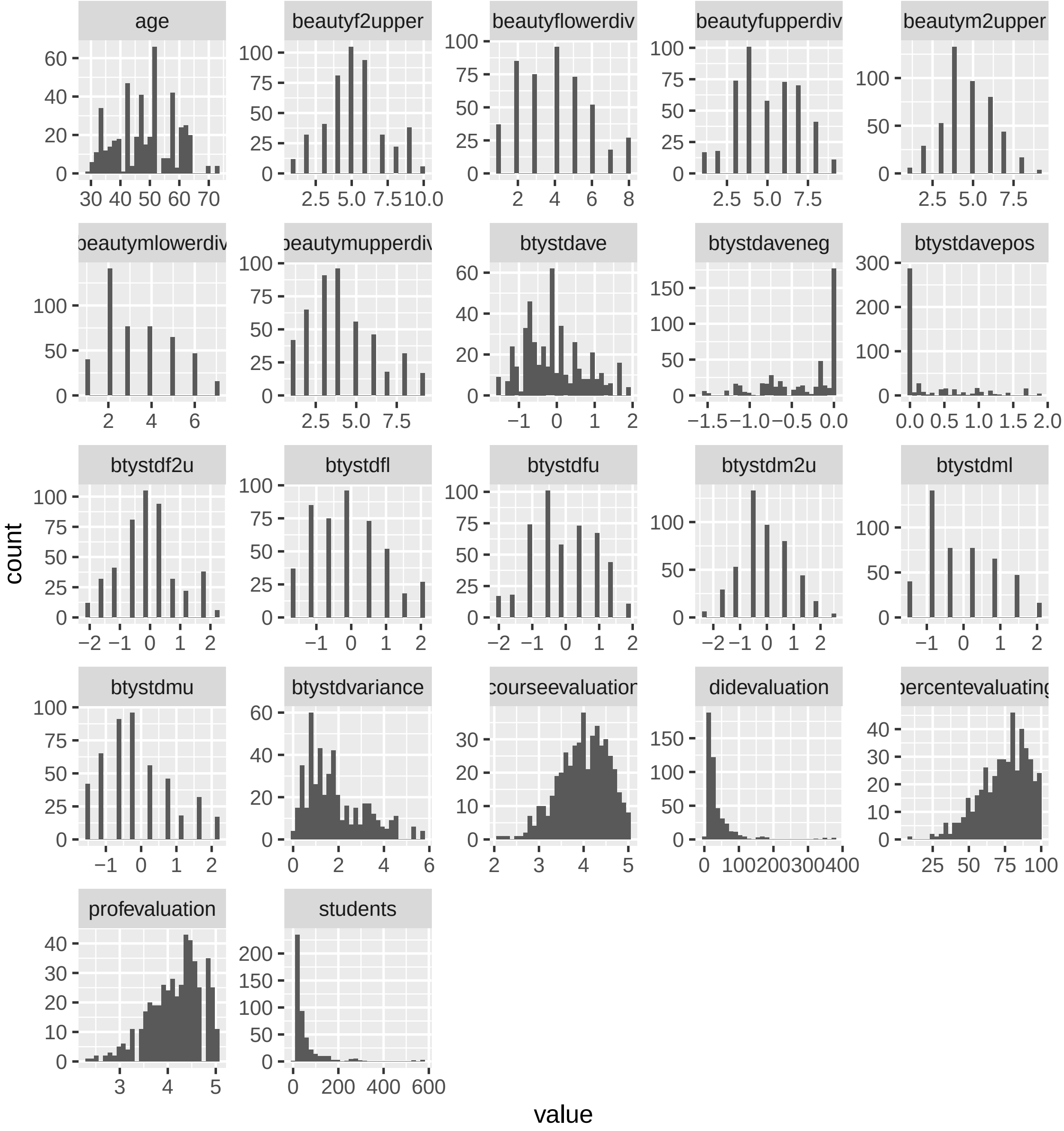

Figure 3: Problem 2: Histograms of all non-indicator variables in the beauty dataset.

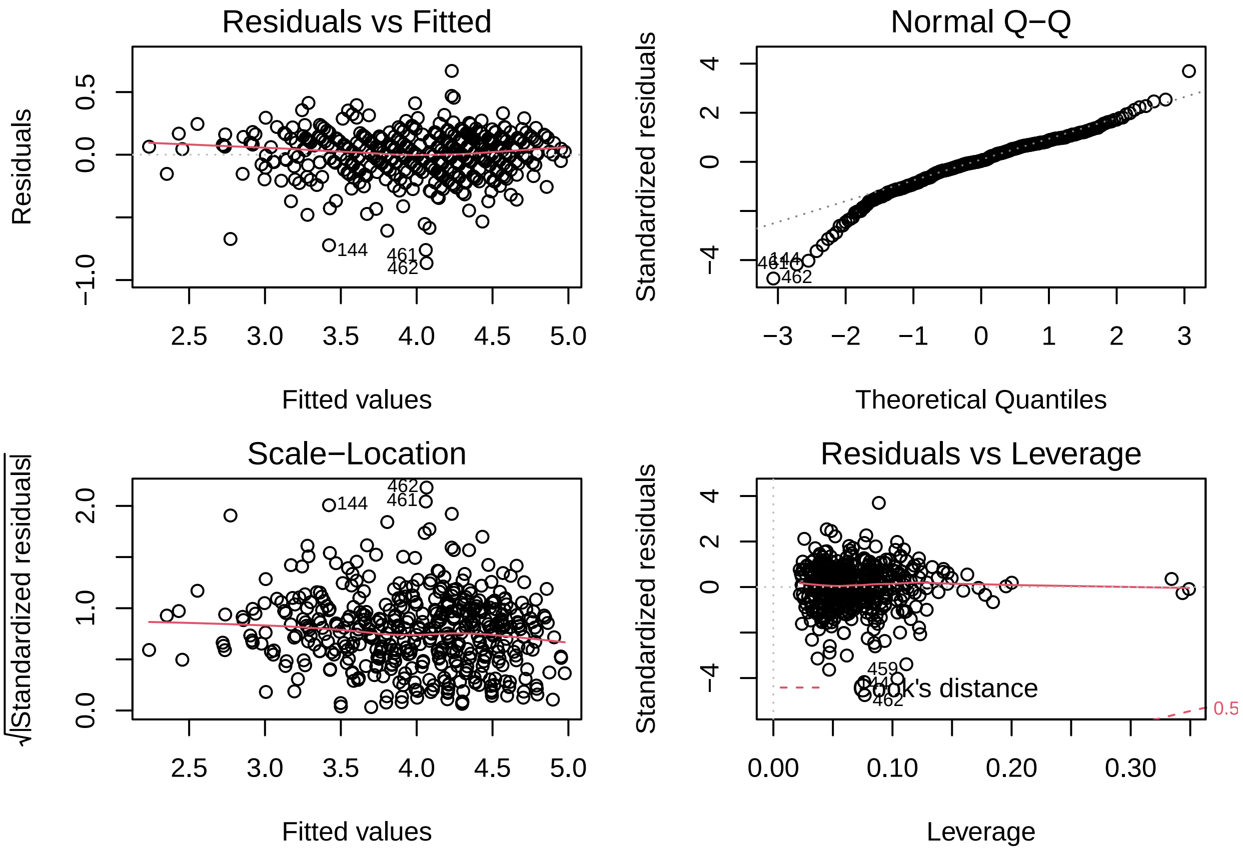

Figure 4: Problem 2b: Diagnostics.

(d) The same issues considered in part (e) of the previous problem apply here. The two reasons to not consider *t*-statistics and VIF's alone are:

- *t*-statistics tell you about whether a variable has a coefficient significantly different from zero, *after* accounting for all of the other predictors. If the variable with a non-significant *t*-statistic is collinear with another variable or group of variables, it is no obvious which (if any!) predictors to eliminate, without help from a subject matter expert.
- The *t*-statistic is being used in a hypothesis test of the hypothesis $H_0 : \beta = 0$. If the *t*-statistic is large, we have sufficient evidence to reject H_0. But if the *t*-statistic is small, it might be *either* because the β really is zero, *or* because we don't have enough evidence (data) yet to rule out $\beta = 0$.