Stat Methods - Homework 7

Justin Reising

December 7, 2018

8.12) A student who used a regression model that included indicator variables was upset when recieving only the following output on the multiple regression printout: "$X^T X$ SINGULAR". What is the likely source of the difficulty.

The source of singularity of the covariance matrix of the predictor variables is most likely from the student using all of the incidences of one or more categorical variables which would create linear dependece of the columns of X matrix. For example, if a predictor variable is categorical and has 5 unique incidences, then the student likely created 5 indicator variables for each of the incidences rather than using 4. The covariance matrix $X^T X$ should be positive semi-definite with non-zero eigenvalues and invertible as variances on the diagonal should not be zero. Otherwise a variable would be constant for every observation.

8.14) In a regression study of factors affecting learning time for a certain task (measured in minutes), gender of a learner was included as a predictor variable (X_2) that was coded $X_2 = 1$ if male and 0 if female. It was found that $b_2 = 22.3$ and $s\{b_2\} = 3.8$. An observer questioned whether the coding scheme for gender is fair becasue it results in a positive coefficent, leading to longer learning times for males than females. Comment.

This comment doesn't really make sense as the interpretation of the coefficient would be slightly different than if it were not an indicator variable. In this case, since X_2 is coded as 1 for male and 0 for female, the coefficent of 22.3 is interpreted as the estimated mean difference in learning time between males and females holding all other variables constant. This means that males' learning time is 22.3 minutes longer than females holding all other variables constant. The model is adjusted for this reference level by the way the indicator variable is coded. Had the researcher switched the coding to 1 for female and 0 for male, we would expect to see a negative coefficient indicating that females' estimated learning time is less than males on average.

8.16) Refer to Grade Point Average. An assistant to the director of admissions conjectured that the predictive power of the model could be imporved by adding information on whether the student had chosen a major field of concentration at the time the application was submitted. Assume that the regression model (8.33) is appropriate, where X_1 is entrance test score and $X_2 = 1$ if the student had indicated a major field of concentration at the time of application and 0 if undecided.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \qquad (8.33)$$

(a) Explain how each regression coefficient is interpreted here.

- β_0- Estimated mean predicted GPA of undeclared majors at the time of application given ACT held constant (0). However, ACT scores can only take on values 1-36, which would indicate that this parameter may only serve as an anchor for the regression function to ensure zero mean error.
- β_1- Estimated mean change in GPA per 1 unit in ACT score of undeclared majors at the time of application.
- β_2- Estimated mean difference in GPA for students who had declared a major field of concentration than students who did not for any given ACT score.

(b) Fit the regression model and state the estimated regression function.

(c) Test whether the X_2 variable can be dropped from the regression model; use $\alpha = 0.01$. State the alternatives, decision rule, and conclusion.

$$H_0 : \beta_2 = 0 \quad H_1 : \beta_2 \neq 0$$

(d) Obtain the residuals for regression model (8.33) and plot them against X_1 and X_2. Is there any evidence in your plot that it would be helpful to include an interaction term in the model?

> Note that we are going to be looking for some sort of inverse relation between X1 and X2 vs the residuals. Wait to see about the indicator variables added to the data from Dr. Lewis.

8.21) In a regression analysis of on-the-job head injuries of warehouse laborers caused by falling objects, Y is a measure of severity of the injury, X_1 is an index reflecting both the weight of the object and the distance it fell, and X_2 and X_3 are indicator variables for the nature of head protection worn at the time of the accident, coded as follows:

Type of Protection	X_2	X_3
Hard Hat	1	0
Bump Cap	0	1
None	0	0

The response function to be used in the study is $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

(a) Develop a response function for each type of protection category.

- $E[Y_{None}] = \beta_0 + \beta_1 X_1$
- $E[Y_{HardHat}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- $E[Y_{BumpCap}] = \beta_0 + \beta_1 X_1 + \beta_3 X_3$

(b) For each of the following questions, specify alternatives H_0 and H_a for the appropriate test:

- (1) With X_1 fixed, does wearing a bump cap reduce the expected severity of injury as compared with wearing no protection?

$$H_0 : \beta_3 = 0 \qquad H_1 : \beta_3 \neq 0$$

Since β_3 is the estimated mean difference in severity than estimated mean severity of no protection with X_1 held constant.

- (2) With X_1 fixed, is the expected severity of injury the same when wearing a hard hat as when wearing a bump cap?

$$H_0 : \beta_2 = \beta_3 \qquad H_1 : \beta_2 \neq \beta_3$$

Since $\beta_2 = \beta_3$ would indicate the estimated mean differences for hard hats and bump caps in injury severity from the estimated mean severity of no protection with X_1 held constant would be the same.

8.22) Refer to tool wear regression model (8.36). Suppose that the indicator variables had been defined as follows: $X_2 = 1$ if tool model M2 and 0 otherwise, $X_3 = 1$ if tool model M3 and 0 otherwise, $X_4 = 1$ if tool model M4 and 0 otherwise. Indicate the meaning of each of the following:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon_i \qquad (8.36)$$

- (1) β_0 - The estimated mean tool wear of tool model "M1" for a given tool speed of 0. In context, a tool speed of 0 may not be interpretable for anything aside from anchoring the regression function (depending on what type of tools are in consideration).
- (2) $\beta_4 - \beta_3$ - The estimated mean difference in tool wear between tool models "M4" and "M3" for any given tool speed.
- (3) β_1 - The estimated mean change in tool wear per 1 unit of tool speed for tool model "M1".

8.27) An analyst wiches to include number of older siblings in family as a predictor variable in a regression analysis of factors affecting maturation in eighth graders. The number of older siblings in the sample observations ranges from 0 to 4. Discuss whether this variable should be placed in the model as an ordinary quantitative variable or by means of 4 0,1 indicator variables.

I would be more concerned about how the analyst is measuring maturity in eighth graders with a numeric variable and the validity of such a subjective measure. However, number of siblings is not a simple discrete variable as the analyst is interested

8.28) Refer to regression model (8.31) for the insurance innovation study. Suppose β_0 were dropped from the model to eliminate the linear dependence in the X matrix so that the model becomes $Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$. What is the meaning here of each of the regression coefficients β_1, β_2, and β_3?

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \qquad (8.31)$$

Although dropping the intercept term β_0 is not recommended for accomodating linear dependence in the X matrix, there are advantages to removing the intercept term in making infernece for the coefficients. In this case, by removing the intercept term we obtain the following expected values for speed of innovation:

- $E[Y_{size}] = \beta_1$ - Since X_1 is not an indicator variable, we would interpret β_1 as the predicted innovation speed when firm size is 0.
- $E[Y_{stock}] = \beta_2$ - Since X_2 is an indicator variable for type of firm (reference stock), we would interpret β_2 as the mean innovation speed for stock firms.
- $E[Y_{mutual}] = \beta_3$ - Since X_3 is an indicator variable for type of firm (reference mutual), we would interpret β_3 as the mean innovation speed for mutual firms.

Also note that removing the intercept term from the model would also have an affect on R^2 in such a way that would inflate the value. Interpreting R^2 can be challenging in this case and I am not sure exactly what the new interpretation could be. Although it would not reflect an increase in goodness of fit. The effect on the definition would be the following:

$$R_0^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i)^2}$$

8.34) In a regression study, three types of banks were involved, namely, commercial, mutual savings, and savings and loan. Consider the following system of indicator variables for type of bank:

Type of Bank	X_2	X_3
Commercial	1	0
Mutual Savings	0	1
Savings & Loan	-1	-1

(a) Develop a first-order linear regression model for relating last year's profit or loss (Y) to size of bank (X_1) and type of bank (X_2, X_3).

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i$$

(b) State the response functions for the three types of banks.

- $E[Y_C] = (\beta_0 + \beta_2) + \beta_1 X_1$
- $E[Y_M] = (\beta_0 + \beta_3) + \beta_1 X_1$
- $E[Y_S] = (\beta_0 - \beta_2 - \beta_3) + \beta_1 X_1$

(c) Interpret each of the following quatities:

- (1) β_2 - Difference of intercept between Commercial banks and Savings & Loan banks in opposite directions.
- (2) β_3 - Difference of intercept between Mutual Savings banks and Savings & Loan banks in opposite directions.
- (3) $-\beta_2 - \beta_3$ - Difference of intercept between Commercial banks and Mutual Savings banks.

11.1) One student remarked to another: "Your residuals show that nonconstancy of error variance is clearly present. Therefore, your regression results are completely invalid."Comment.

This is not completely true. Constant error variance is not a single indicator of validity of a model per say. If the response varable is Poisson distributed with a mean that increases as a predictor variabele increases, then the response cannot have constant variance at all levels of the predictor since the variance of a Poisson variable equals the mean, which is increasing with the predictor.

11.2) An analyst suggested: "One nice thing about robust regression is that you need not worry about outliers and influential observations." Comment.

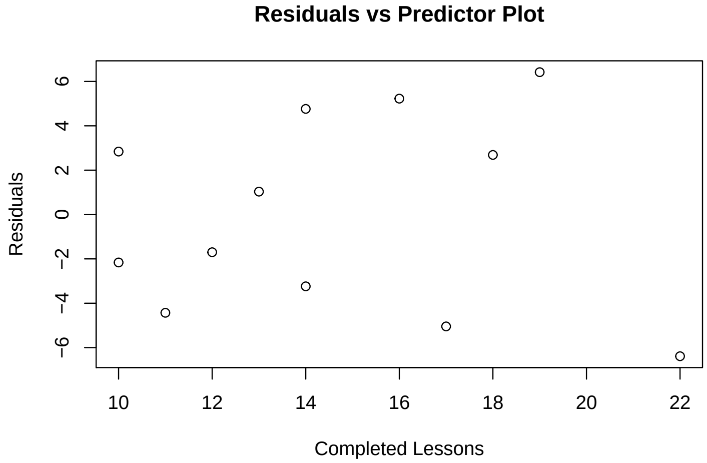
While robust regression procedures dampen the influence of outlying cases as compared to ordinary least squares estimation, it is not a "sure thing" in worrying about outliers and influential observations. It could be the case that removing such influential observaitons from the model could have a negative affect on the adequacy of the model and robust regression procedures do not help matters either. Typically, these influential observations should be examined in greater detail to determine how to handle them in regard to consideration for removal, but it also can depend on the size of your dataset and how much of an influence an observation may actually be. In fact, robsust regression can confirm if an influential observation really has an effect on the ordinary least squares estimate.

11.6) Computer Assisted Learning. Data from a study of computer-assisted learning by 12 students, showing the total number of responses in completing a lesson (X) and the coost of computer time (Y in cents) follow:

Table 3: Computer Assisted Learning

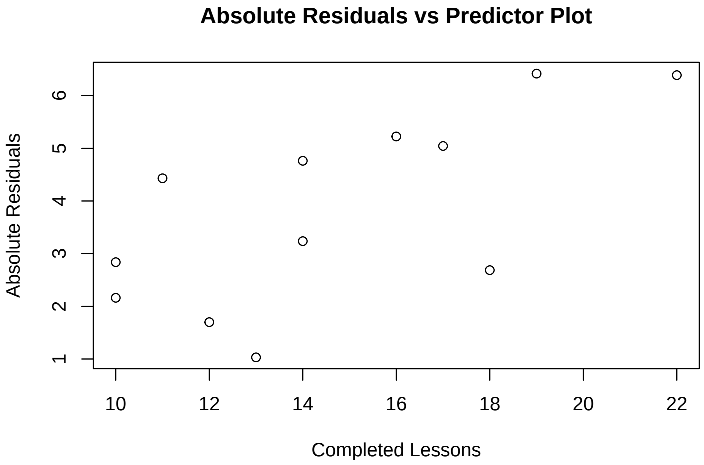
Y	77	70	85	50	62	70	55	63	88	57	81	51
X	16	14	22	10	14	17	10	13	19	12	18	11

(a) Fit a linear regression function by ordinary least squares, obtain the residuals, and plot the residuals against X. What does the residual plot suggest?



The plot above shows that there may be evidence of non-constant variance and suggests that we may want to consider regressing the absolute residuals against X.

(c) Plot the absolute values of the residuals against X. What does this plot suggest about the relation between the standard deviation of the error term and X?



This shows that there is a positive linear relation between the predictor variable X and the absolute residuals and verify's the fan pattern in the previous plot. This suggests that there a linear relation between the error standard deviation and X may be reasonable.

(d) Estimate the standard deviation function by regressing absolute values of the residuals against X, and then calculate the estimated weights for each case using (11.16a). Which case recieve the largest weight? Which case recieves the smallest weight?

$$w_i = \frac{1}{(\hat{s}_i)^2} \qquad (11.16a)$$

- Estimated standard deviation function (OLS): $\hat{s} = -.905 + 0.32X$
- Max Weight obtained by observation 4.
- Min Weight obtained by observation 3.

(e) Using the estimated weights, obtain the weighted least squares estimates of β_0 and β_1. Are these estimates similar to the ones obtained with the ordinary least squares in part (a)?

Table 4: OLS Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.4727	5.5162	3.5301	0.0054
X	3.2689	0.3651	8.9546	0.0000

Table 5: WLS Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.3006	4.8277	3.5836	0.005
X	3.4211	0.3703	9.2385	0.000

In tables 4 and 5 we can see that he estimates for β_1 are fairly similar where the estimates for β_0 are al ittle more than 2 units more in the OLS model than in the WLS model.

(f) Compare the estimated standard deviation of the weighted least squares estimates b_{w0} and b_{w1} in part (c) with those for the ordinary least squares in part (a). What do you find?

- Estimated standard deviation function (WLS): $\hat{s} = -1.5711 + 0.3646X$

  b_{w0} appears to be about 0.8 less than the estimate with OLS and b_{w1} is approximately the same differing only by 0.04.

(Additional) Compute the bootstrap confidence interval for β_1 using both the percentile method and reflection method. Also, state which resampling method you choose and why (fixed X or random X). Compare the mean and standard deviation of the the empirical distribution to that found using weighted least squares regression.

- Weighted Least Squares CI:

```
alpha<-.05
b1 <- coef(comp.lm.wls)[2]
# getting a 95% confidence interval for the true slope beta_1 :
confint(comp.lm.wls,"X",level=.95)
```

```
##      2.5 %   97.5 %
## X 2.596004 4.246208
```

```
lower <- confint(comp.lm.wls,"X",level=.95)[1]
upper <- confint(comp.lm.wls,"X",level=.95)[2]
print(paste(100*(1-alpha), "percent CI for slope:", round(lower,4), round(upper,4)))
```

```
## [1] "95 percent CI for slope: 2.596 4.2462"
```

- Resampling Method: Random X - Since there is evidence of non-constant error variance in OLS Model.

```
### Random X resampling:
B = 1000
b1.star.vec = rep(0,times=B)
for (i in 1:B) {
boot.samp <- computer[sample(1:(nrow(computer)), replace=T),]
comp.reg.boot <- lm(Y ~ X, data=boot.samp)
abs.res.boot <- abs(resid(comp.reg.boot))
comp.reg2.boot <- lm(abs.res.boot ~ X, data=boot.samp)
weight.vec.boot <- 1/((fitted(comp.reg2.boot))^2);
comp.wls.reg.boot <- lm(Y ~ X, weights = weight.vec.boot, data=boot.samp)
b1.star.vec[i] <- coef(comp.wls.reg.boot)[2]
}
```

- Percentile Method CI:

```
# Percentile-method 95% CI for beta_1:

L.p <- quantile(b1.star.vec, alpha/2)
U.p <- quantile(b1.star.vec, 1-alpha/2)
print(paste("Percentile-method 95% CI for beta_1:", round(L.p,4), round(U.p,4)) )
```

```
## [1] "Percentile-method 95% CI for beta_1: 2.6503 4.289"
```

- Reflection Method CI:

```
# Reflection-method 95% CI for beta_1:

d1 <- b1 - quantile(b1.star.vec, alpha/2)
d2 <- quantile(b1.star.vec, 1-alpha/2) - b1
L.r <- b1 - d2
```

```
U.r <- b1 + d1
print(paste("Reflection-method 95% CI for beta_1:", round(L.r,4), round(U.r,4)) )
```

```
## [1] "Reflection-method 95% CI for beta_1: 2.5532 4.1919"
```

- Compare Standard Deviations

```
# Comparing the Standard Deviations
summary(comp.lm.wls)
```

```
##
## Call:
## lm(formula = Y ~ X, weights = w_i)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48741 -0.96167 -0.04198  1.10930  1.50265
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)  17.3006     4.8277   3.584    0.00498 **
## X             3.4211     0.3703   9.238 0.00000327 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.159 on 10 degrees of freedom
## Multiple R-squared:  0.8951, Adjusted R-squared:  0.8846
## F-statistic: 85.35 on 1 and 10 DF,  p-value: 0.000003269
```

```
sd<- .3703
sd.star<-sd(b1.star.vec)
sd
```

```
## [1] 0.3703
```

```
sd.star
```

```
## [1] 0.4444024
```

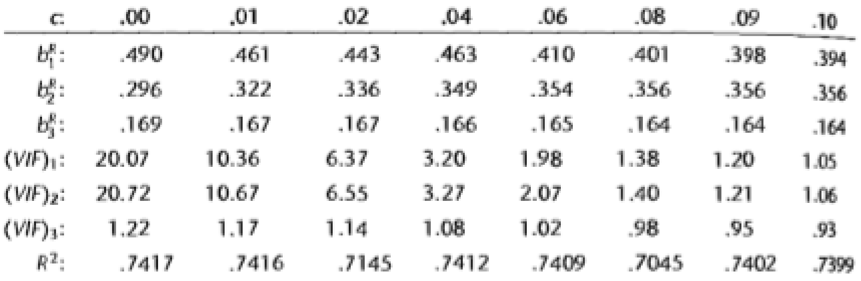
```
b1_wls<-3.4211
b1_wls
```

```
## [1] 3.4211
```

```
mean(b1.star.vec)
```

```
## [1] 3.448279
```

Note that the mean estimates of β_1 of the empirical and the WLS model are practically the same, although the standard deviations are slightly different where the emperical distribution has a slightly higher standard deviation.



c:	.00	.01	.02	.04	.06	.08	.09	.10
b_1^R:	.490	.461	.443	.463	.410	.401	.398	.394
b_2^R:	.296	.322	.336	.349	.354	.356	.356	.356
b_3^R:	.169	.167	.167	.166	.165	.164	.164	.164
$(VIF)_1$:	20.07	10.36	6.37	3.20	1.98	1.38	1.20	1.05
$(VIF)_2$:	20.72	10.67	6.55	3.27	2.07	1.40	1.21	1.06
$(VIF)_3$:	1.22	1.17	1.14	1.08	1.02	.98	.95	.93
R^2:	.7417	.7416	.7145	.7412	.7409	.7045	.7402	.7399

Figure 1: Ridge Regression Coefficients

11.9) Refer to Cosmetics Sales. Given above in Figure 1 are the estimated ridge standardized regression coeeficients, the variance inflation factors, and R^2 for selected biasing constants c.

(a) Fit an ordinary ;east squares regression of Y against X_1, X_2, and X_3. Calculate the VIFs. What do these tell you?

```
cos.lm.ols<- lm(Sales~X1+X2+X3)
vif(cos.lm.ols)
```

```
##        X1        X2        X3
## 20.072031 20.716101  1.217973
```

With the VIF values approximately 20 for X_1 and X_2, there is evidence of significant multicolinearity in these predictor variables.

(b) Fit the ridge regression of Y against X_1, X_2, and X_3 using a biasing constant of $\lambda = 0.1$. Write the fitted regression equation.

$$Y_i = 1.06 + 0.91X_1 + 0.69X_2 + 0.67X_3$$

(c) How do the SSE's for the two models compare? What about the R^2 and VIF values?

```
ols.sum<-summary(cos.lm.ols)
#ridge.sum<-summary(cosmetic.ridge)
# Transforming the Variables Using Correlation Transformation
Sales2<-(1/(sqrt(length(Sales)-1)))*((Sales-mean(Sales))/sd(Sales))
X1_2<-(1/(sqrt(length(Sales)-1)))*((X1-mean(X1))/sd(X1))
X2_2<-(1/(sqrt(length(Sales)-1)))*((X2-mean(X2))/sd(X2))
X3_2<-(1/(sqrt(length(Sales)-1)))*((X3-mean(X3))/sd(X3))

cosmetic.reg.trans<- lm(Sales2 ~ X1_2 + X2_2 + X3_2)
anova(cosmetic.reg.trans)
```

```
## Analysis of Variance Table
##
## Response: Sales2
##           Df  Sum Sq Mean Sq  F value             Pr(>F)
## X1_2       1 0.70852 0.70852 109.7054 0.0000000000004994 ***
## X2_2       1 0.00983 0.00983   1.5215            0.22459
## X3_2       1 0.02332 0.02332   3.6113            0.06461 .
## Residuals 40 0.25833 0.00646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
sse.ols<-0.2583

sd.org.y<-sd(Sales)
sd.org.X1<-sd(X1)
sd.org.X2<-sd(X2)
sd.org.X3<-sd(X3)

b1.trans.X1<-0.9072*(sd.org.X1/sd(Sales))
b2.trans.X2<-0.6850*(sd.org.X2/sd(Sales))
b3.trans.X3<--0.6710*(sd.org.X3/sd(Sales))

yhat.trans<-b1.trans.X1*X1_2+b2.trans.X2*X2_2+ b3.trans.X3*X3_2
sse.ridge<-sum((Sales2-yhat.trans)^2)
sse.ridge
```

```
## [1] 0.3714421
```

```r
sse.ols
```

```
## [1] 0.2583
```

- R^2 - (OLS) 0.7223 - (Ridge) 0.73470

- Ridge VIFs X1 X2 X3 k=0.01 10.35846 10.6723 1.17124

  Ridge Regression reduced the VIF values for X_1 and X_2 by a factor of about 2 bringing the threat of multicolinearity down to a semi-acceptable level. The R^2 values remained approximately the same, and the SSE of the Ridge model is slightly higher due to the bias introduced.

11.12)

11.19)