UNIVERSITY OF TORONTO

Faculty of Arts and Science

DECEMBER EXAMINATIONS 2009
STA 302 H1F / STA 1001 HF

Duration - 3 hours

Aids Allowed: Calculator

LAST NAME: SOLUTIONS **FIRST NAME:** ____________

STUDENT NUMBER: ____________

- There are 23 pages including this page.
- The last page is a table of formulae that may be useful. For all questions you can assume that the results on the formula page are known.
- A table of the t distribution can be found on page 19 and tables of the F distribution can be found on pages 20, 21 and 22.
- Total marks: 90

1ab	1cde	2abcd	2ef	3abcde	3fg	3h(i,ii,iii)

3h(iv,v)	4ab	4cd	4e	5	6a	6bc

1. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n$$

where the ϵ_i are independent and identically distributed $N(0, \sigma^2)$ random variables. Assume that the X_i are not random. Let b_0 and b_1 be the least squares estimates of β_0 and β_1 respectively.

(a) (3 marks) Explain the method of least squares as used in simple linear regression. Use language that someone who has studied no statistics can understand.

Find estimates of β_0 and β_1 that minimize the sum of the squares of the vertical distances from the data points to the fitted line.

(b) (5 marks) Derive the formula for b_1, the least squares estimate of β_1. (Your answer should be one of the expressions on the formula sheet. You may assume that the formula for b_0 is known.)

Minimize

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Take derivative with respect to β_1:

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i)$$

Setting the derivative equal to 0 gives:

$$\sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 = b_0 n \overline{X} + b_1 \sum_{i=1}^{n} X_i^2$$

and since $b_0 = \overline{Y} - b_1 \overline{X}$,

$$\sum_{i=1}^{n} X_i Y_i = n \overline{X}\overline{Y} - n b_1 \overline{X}^2 + b_1 \sum_{i=1}^{n} X_i^2$$

so

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n \overline{X}\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n \overline{X}^2}$$

(Question 1 continued.)

(c) (4 marks) $\hat{Y}_i$ is the value of Y for the ith observation estimated from the fitted regression line. Show that $\sum_{i=1}^n \hat{Y}_i = \sum_{i=1}^n Y_i$.

$$
\begin{aligned}
\sum_{i=1}^{n} \hat{Y}_1 &= \sum_{i=1}^{n} (b_0 + b_1 X_i) \\
&= \sum_{i=1}^{n} (\overline{Y} - b_1 \overline{X} + b_1 X_i) \\
&= n\overline{Y} - b_1 n\overline{X} + b_1 n\overline{X} \\
&= \sum_{i=1}^{n} Y_i
\end{aligned}
$$

(d) (2 marks) The formula for h_{ij} is on the formula sheet. We are usually most interested in h_{ii} ($i = 1, \ldots, n$), the values of h_{ij} when $i = j$. What do the h_{ii} measure? Why are they of interest?

h_{ii} measures how far a point is from the rest of the points in the X direction. High h_{ii} means the ith point has high leverage and may be influential.

(e) (2 marks) Show that $\sum_{i=1}^n h_{ii} = 2$.

$$
\begin{aligned}
\sum_{i=1}^{n} h_{ii} &= \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(X_i - \overline{X})^2}{S_{XX}} \right) \\
&= 1 + \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{S_{XX}} \\
&= 2
\end{aligned}
$$

2. A multiple linear regression model with dependent variable Y and k explanatory variables is fit to n observations $(X_{i1}, X_{i2}, \ldots, X_{ik}, Y_i)$, $i = 1, \ldots, n$. You may assume that the X's are not random.

   (a) (3 marks) State the multiple regression model in matrix terms, defining all matrices and vectors.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{pmatrix}$$

   (b) (5 marks) State all the assumptions of the multiple regression model that are necessary to estimate and carry out inference on the model parameters.

The model is appropriate.
$E(\boldsymbol{\epsilon}) = \mathbf{0}$
$Var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$ *(This covers constant variance and uncorrelated errors.)*
$\boldsymbol{\epsilon}$ *has a multivariate Normal distribution*

   (c) (2 marks) Which assumption is most critical? Why?

The model is appropriate. If not the correct model, everything else is wrong.

   (d) (2 marks) Which assumption is least critical? Why?

Normality. Estimation of parameters is still OK without it and most inferences are still approximately correct because of the Central Limit Theorem.

(Question 2 continued.)

(e) (3 marks) What is the variance-covariance matrix of the vector of the fitted values of Y?

$$\begin{aligned} Cov(\hat{\mathbf{Y}}) &= Cov(\mathbf{HY}) \quad where\ \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ Cov(\mathbf{Y})\ \left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ \sigma^2\mathbf{I}\ \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \sigma^2\mathbf{H} \end{aligned}$$

(f) (2 marks) Show that the vector of least squares estimators $\mathbf{b}$ is unbiased for the vector of model parameters $\boldsymbol{\beta}$.

$$\begin{aligned} E(\mathbf{b}) &= E\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ E(\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

3. A company that publishes a newspaper in a mid-size American city wants to investigate the feasibility of introducing a Sunday edition of the paper. The current circulation (the average number of newspapers sold per day) of the company's weekday newspaper is 210,000. The goal of this analysis is to predict the Sunday circulation of a newspaper with a weekday circulation of 210,000.

The data are circulations of 89 U.S. newspapers that publish both weekday and Sunday editions.

Analysis was carried out on the natural logarithms of the circulations.

Some output from SAS is below. `logSunCirc` is the natural logarithm of the Sunday circulation and `logWkdayCirc` is the natural logarithm of the weekday circulation.

Descriptive Statistics

Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	89.00000	1.00000	89.00000	0	0
logWkdayCirc	1101.17108	12.37271	13651	0.30582	0.55301
logSunCirc	1126.23888	12.65437	14281	0.33017	0.57460

The REG Procedure
Dependent Variable: logSunCirc

Number of Observations Read	89
Number of Observations Used	89

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	26.20543	26.20543	800.19	<.0001
Error	87	2.84916	0.03275		
Corrected Total	88	29.05458			

Root MSE	0.18097	R-Square	0.9019
Dependent Mean	12.65437	Adj R-Sq	0.9008
Coeff Var	1.43007		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.44511	0.43204	1.03	0.3057
logWkdayCirc	1	0.98679	0.03488	28.29	<.0001

Questions about this output begin on the next page.

(Question 3 continued.)

(a) (1 mark) What are the null and alternative hypotheses for the test with test statistic 800.19?

$H_0:\ \beta_1 = 0$ *versus* $H_a:\ \beta_1 \neq 0$

(b) (1 mark) What percent of variability in the dependent variable is explained by its linear relationship with the independent variable?

$R^2 = 90.19\%$

(c) (3 marks) Use the Bonferroni method to find simultaneous 90% confidence intervals for the slope and the intercept.

$t_{87,.10/2/2} = 2.0$ *(approximating with 60 df)*
Intercept: $.44511 \pm 2(.43204) = (-.419, 1.31)$
Slope: $.98679 \pm 2(.03488) = (.917, 1.056)$

(d) (2 marks) What does it mean for the confidence intervals in part (c) to be "simultaneous"?

The probability that **both** *capture their respective parameters is* **at least** *90%.*

(e) (2 marks) Explain how to interpret the estimated slope in a practical way. Your answer should be in terms of the original circulations and not the log transformations of them.

A change in weekday circulation by a factor of k results in an average change in Sunday circulation by a factor of $k^{.98679}$.

(Question 3 continued.)

(f) (2 marks) In order to satisfy the usual regression model assumptions, it was necessary to take the log transform of both Sunday circulation and weekday circulation. Describe the features of the plot of Sunday circulation versus weekday circulation (that is, the scatterplot of the variables before transformation) that indicate that the log transformation of both of the variables is necessary?

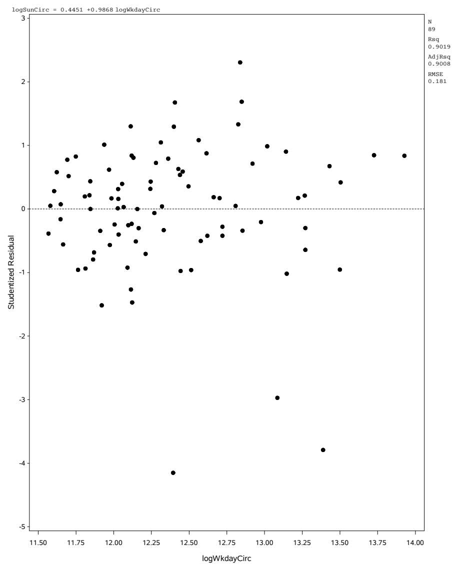
Linear
Increasing variance

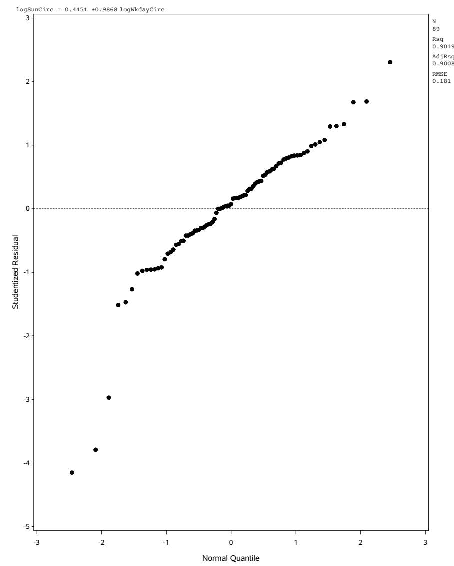
(g) (5 marks) Calculate a 95% interval estimate of the Sunday circulation for the newspaper that is considering adding a Sunday newspaper. (Recall that its weekday circulation is 210,000.) (Note that you'll first need to calculate the interval for the log of the circulation, and then back-transform it for an interval estimate of the circulation.)

$\widehat{\texttt{logSunCirc}} = .44511 + .98679 \log(210000) = 12.538$
$t_{.025,87} \doteq 2.00$
Prediction interval: $12.538 \pm 2.00(.18097)\sqrt{1 + \frac{1}{89} + \frac{(\log(210000) - 12.27271)^2}{88(.30582)}} = (12.174, 12.902)$
Interval for predication of Sunday circulation: $(e^{12.174}, e^{12.902}) = (193687, 401114)$

(Question 3 continued.)

(h) The plots below are a plot of the studentized residuals versus the explanatory variable and a normal quantile plot of the studentized residuals for the regression whose output is on page 6.





i. (1 mark) You are not given the plot of the studentized residuals versus the predicted values. Describe what it would look like.

Just like the first plot except the horizontal scale would be different.

ii. (4 marks) What are you looking for in the plot of the studentized residuals versus the explanatory variable? What do you conclude?

No curvature – not a problem
Constant variance – not a problem
Influential points – not a problem
Outliers – there are a few large, negative residuals

iii. (2 marks) What additional information do you learn from the normal quantile plot of the studentized residuals?

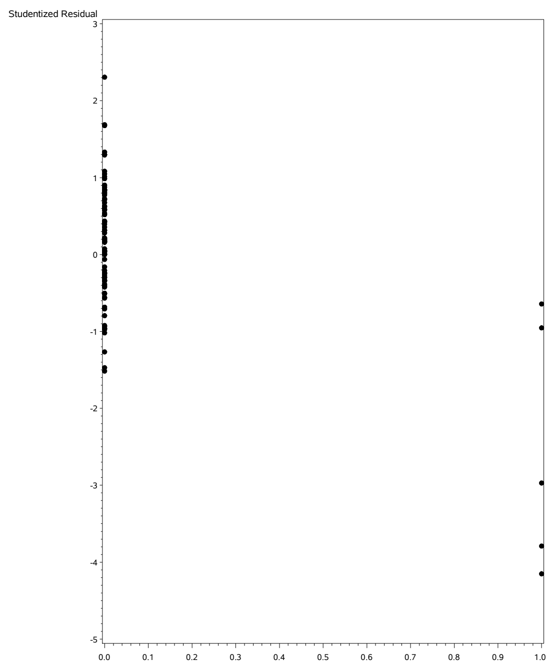
The only problem with normality is the existence of the outliers.

(Question 3 continued.)

iv. (2 marks) Studentized rather than raw residuals were plotted. (The “raw” residuals are $e_i = Y_i - \hat{Y}_i$.) What are the advantages of looking at plots of studentized residuals rather than raw residuals?

Variances are constant.
Standardization makes outliers easier to identify.

v. (2 marks) Some of the newspapers in the dataset are tabloids. Below is a plot of the studentized residuals versus an indicator variable which is 1 if the newspaper is a tabloid and 0 otherwise. What additional information do you learn from this plot?



Tabloids all have negative residuals and include the outliers. We should include the dummy variable in the model.

4. Which aspects of a professional golfer's play are most important in determining the amount of prize money he will earn? In order to answer this question, data on the top 196 professional golfers in 2006 were collected. The aspects of a golfer's play we will consider are listed below. For each aspect, it is noted whether a high or low value indicates that the golfer is performing well.

   - `daccuracy`: Driving Accuracy is the percent of time the golfer hit the fairway from the tee. High values are good.
   - `gir`: Greens in Regulation is the percent of time the golfer hit the green in the number of shots allotted for this. High values are good.
   - `puttavg`: Putting Average is the average number of putts to score on holes where the green is hit in regulation. Low values are good.
   - `birdies`: The percent of time that the golfer makes a birdie after hitting the green in regulation. High values are good.
   - `sandsaves`: The percent of time a golfer was able to get out of a sand bunker. High values are good.
   - `scrambling`: The percent of time a golfer misses the green in regulation but recovers. High values are good.
   - `nputts`: The average number of putts per round. Low values are good.

   The analysis has been carried out using the natural logarithm of prize money (`logprizemoney`) as the dependent variable. Some output from SAS is below.

```
                    The REG Procedure
Number of Observations Read                          196
Number of Observations Used                          193
Number of Observations with Missing Values             3
```

Correlation

Variable	daccuracy	gir	puttavg	birdies
daccuracy	1.0000	0.4114	-0.0176	-0.2637
gir	0.4114	1.0000	0.0710	0.0130
puttavg	-0.0176	0.0710	1.0000	-0.7662
birdies	-0.2637	0.0130	-0.7662	1.0000
sandsaves	0.0348	-0.0804	-0.2651	0.1305
scrambling	0.3859	0.1830	-0.1939	-0.0335
nputts	0.0702	0.4935	0.7941	-0.5060
logprizemoney	0.1667	0.4936	-0.4215	0.4590

Correlation

Variable	sandsaves	scrambling	nputts	logprizemoney
daccuracy	0.0348	0.3859	0.0702	0.1667
gir	-0.0804	0.1830	0.4935	0.4936
puttavg	-0.2651	-0.1939	0.7941	-0.4215
birdies	0.1305	-0.0335	-0.5060	0.4590
sandsaves	1.0000	0.5058	-0.4207	0.2457
scrambling	0.5058	1.0000	-0.4079	0.3519
nputts	-0.4207	-0.4079	1.0000	-0.1745
logprizemoney	0.2457	0.3519	-0.1745	1.0000

(SAS output for this question continues on the next page.)

(Question 4 continued.)

The REG Procedure
Model: MODEL1
Dependent Variable: logprizemoney

Number of Observations Read	196
Number of Observations Used	193
Number of Observations with Missing Values	3

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	97.41905	13.91701	31.31	<.0001
Error	185	82.23911	0.44454		
Corrected Total	192	179.65817			

Root MSE	0.66674	R-Square	0.5422
Dependent Mean	10.39148	Adj R-Sq	XXX
Coeff Var	6.41617		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.61671	7.83849	0.08	0.9374	0
daccuracy	1	-0.00413	0.01186	-0.35	0.7280	1.78632
gir	1	0.19557	0.04440	4.40	<.0001	6.27168
puttavg	1	-1.01815	6.96608	-0.15	0.8840	12.86093
birdies	1	0.15287	0.04092	3.74	0.0002	3.48775
sandsaves	1	0.01621	0.00997	1.63	0.1057	1.47982
scrambling	1	0.04953	0.03199	1.55	0.1232	4.30293
nputts	1	-0.30758	0.47911	-0.64	0.5217	19.45000

(a) (3 marks) Calculate adjusted R^2. In multiple regression, why is it preferred over R^2?

$1 - 192(0.44454/179.65817) = 0.525$

R^2 always increases wtih additional predictors but adjusted R^2 only goes up if the predictors are useful and MSE goes down.

(b) (2 marks) Give a practical interpretation for the coefficient of `birdies`.

For golfers with every other aspect of the game the same, a 1% increase in making birdies increases prize money by a factor of $e^{.15287}$.

(Question 4 continued.)

(c) (2 marks) The p-value for the t-test for the coefficient of `birdies` is 0.0002. What do you conclude from this?

Given the other variables are in the model, we have strong evidence that the coefficient of birdies is not 0.

(d) (2 marks) Give 2 indications from the SAS output above that there are problems with multicollinearity.

Some VIFs are greater than 10.
There are correlations between some predictors from the table on p. 11.

(e) The t-tests for the coefficients of 5 of the predictor variables (`daccuracy`, `puttavg`, `sandsaves`, `scrambling` and `nputts`) have high p-values. These 5 predictors were removed from the model and the data were re-fit to the reduced model giving the following SAS output:

```
                    The REG Procedure
                      Model: MODEL1
             Dependent Variable: logprizemoney

    Number of Observations Read                        196
    Number of Observations Used                        193
    Number of Observations with Missing Values           3
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	80.58561	40.29280	77.27	<.0001
Error	190	99.07256	0.52143		
Corrected Total	192	179.65817			

Root MSE	0.72210	R-Square	0.4485
Dependent Mean	10.39148	Adj R-Sq	0.4427
Coeff Var	6.94900		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-6.72601	1.42201	-4.73	<.0001
gir	1	0.17385	0.01920	9.05	<.0001
birdies	1	0.19940	0.02373	8.40	<.0001

(Questions continue on the next page.)

i. (2 marks) Was it a good idea to remove the 5 predictors with high p-values and fit this reduced model? Why or why not?

No, not all at once. The t-tests assume that all other variables are in the model; removing one may result in coefficients of other variables becoming significant.

ii. (4 marks) Carry out an hypothesis to test simultaneously whether all of the variables removed from the original model have coefficients equal to 0.

Partial F-test
Test statistic: $F_{obs} = \frac{(97.419-80.586)/5}{.44454} = 7.57$
Under H_0 *that the 5 coefficients are 0, this is an observation from an* $F(5, 185)$ *distribution. Approximating with the* $F(5, 120)$ *distribution, we can conclude* $p < 0.001$.
So we have strong evidence that at least one of the coefficients is not 0.

5. In Assignment 1, we examined whether there is a relationship between an NFL kicker's field-goal percentage one year and the previous year. For each of 19 kickers, we have data for four consecutive years. The analysis of these data below has as the dependent variable the percentage of field goals scored in one year (`FG`) and as independent variables: the percentage of field goals scored in the previous year (`prevFG`) and 18 indicator variables for the football players. For example, the indicator variable `AV` is 1 if the observation is for the kicker with initials A.V. and 0 if the observation is for another kicker. The regression below fits 19 separate lines, one for each player. In this regression, the lines are parallel.

```
                    The REG Procedure
                 Dependent Variable: FG

          Number of Observations Read         76
          Number of Observations Used         76

                  Analysis of Variance
                           Sum of           Mean
Source             DF     Squares         Square    F Value    Pr > F
Model              19  2339.66699      123.14037       3.19    0.0004
Error              56  2160.95656       38.58851
Corrected Total    75  4500.62355

     Root MSE              6.21197    R-Square     0.5199
     Dependent Mean       82.25921    Adj R-Sq     0.3569
     Coeff Var             7.55170

                 Parameter Estimates
                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|
Intercept     1      116.31347        9.32238      12.48      <.0001
prevFG        1       -0.50370        0.11276      -4.47      <.0001
AV            1       10.37368        4.45141       2.33      0.0234
DA            1        5.72740        4.41591       1.30      0.2000
JE            1        7.35703        4.39774       1.67      0.0999
JaH           1       12.49090        4.47697       2.79      0.0072
JR            1        2.07814        4.41828       0.47      0.6399
JW            1       12.68387        4.44053       2.86      0.0060
JC            1        4.39629        4.40068       1.00      0.3221
JoH           1        1.88722        4.39253       0.43      0.6691
KB            1       -2.98610        4.40553      -0.68      0.5007
MS            1       19.10997        4.51993       4.23      <.0001
MV            1       15.26923        4.49769       3.39      0.0013
NR            1        3.75369        4.42014       0.85      0.3994
OM            1       -2.66278        4.39253      -0.61      0.5468
PD            1       13.92609        4.46388       3.12      0.0029
RiL           1        5.50629        4.39670       1.25      0.2156
RyL           1        8.14221        4.42371       1.84      0.0710
SJ            1        6.39740        4.40287       1.45      0.1518
SG            1       12.50869        4.43971       2.82      0.0067
```

Questions begin on the next page.

(Question 5 continued.)

(a) (2 marks) What is the fitted regression line for the kicker with initials S.G.?

$\hat{\texttt{FG}} = 116.31 + 12.51 - 0.5037\texttt{prevFG}$

(b) (2 marks) If 19 indicator variables had been included in the model, one for each of the 19 kickers, SAS would have deleted one. Why?

The columns of the $\mathbf{X}$ *matrix would be linearly dependent as the 19 columns for the indicator variables would sum to* $\mathbf{1}$*, the first column of the* $\mathbf{X}$ *matrix.*

(c) (1 mark) Before fitting the parallel lines model as above, it should first be checked whether or not the data are consistent with parallel lines for each kicker. How should the model fit above be adjusted so that a different slope (and intercept) is estimated for each kicker?

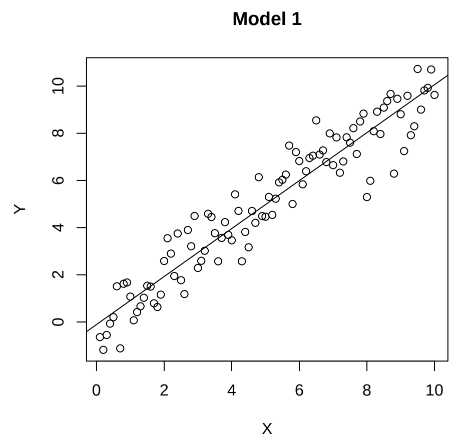
Add 18 interaction terms: the dummy variable for each kicker times `prevFG`*.*

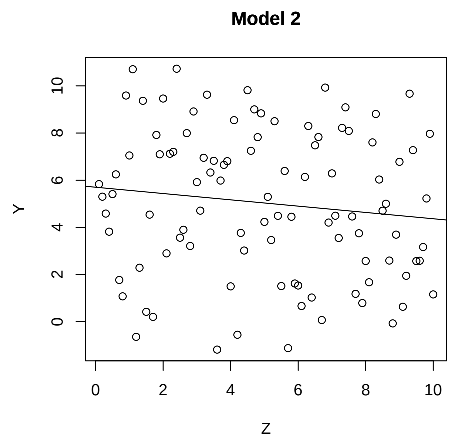
(d) (3 marks) A statistical test can be carried out on the model in part (c) to determine whether it was reasonable to model the 19 lines as parallel. Indicate the type of test that is appropriate, the null and alternative hypotheses, and the distribution of the test statistic under the null hypothesis.

Partial F-test
$H_0: \beta_{20} = \beta_{21} = \cdots = \beta_{37} = 0$ *where* $\beta_{20}, \ldots, \beta_{37}$ *are the coefficients of the interaction terms in part (c)*
H_a *: at least one of* $\beta_{20}, \ldots, \beta_{37}$ *is not 0*
Under H_0*, the test statistic would be an observation from an* $F(18, 56 - 18 = 38)$ *distribution.*

6. The following questions require short answers.

(a) (3 marks) Two alternative straight line regression models have been proposed for Y. In the first model, Y is a linear function of X, while in the second model Y is a linear function of Z. The plots below show scatterplots with the fitted regression lines of Y versus X (first plot) and Y versus Z (second plot).





Which one of the following statements is true? Give a detailed reason to support your choice.

i. SSE for model 1 is greater for SSE for model 2, while SSR for model 1 is greater than SSR for model 2.
ii. SSE for model 1 is less than SSE for model 2, while SSR for model 1 is less than SSR for model 2.
iii. SSE for model 1 is greater than SSE for model 2, while SSR for model 1 is less than SSR for model 2.
iv. SSE for model 1 is less than SSE for model 2, while SSR for model 1 is greater than SSR for model 2.

iv.
SSTO is the same for both.
$SSR_1 > SSR_2$ *since the amount of variability in the observations explained by the regression line is greater for model 1*
Since $SSR + SSE = SSTO$, $SSE_1 < SSE_2$

(Question 6 continued.)

(b) (3 marks) Suppose a multiple regression model with 5 predictor variables is fit to some data. The analysis of variance F-test is statistically significant (p-value< 0.01) but the t-tests for the coefficients of the predictor variables are all not statistically significant (all 5 p-values are > 0.10). What do you conclude? Explain.

From the F-test, we conclude that not all of $\beta_1, \ldots, \beta_5$ are 0.
There must be some relationships among the X's since the p-values for the t-tests lead us to conclude that each predictor has a coefficient not statistically significantly different from 0, given the others are in the model. (We don't know which predictors are important.)

(c) (1 mark) Explain the purpose of using centering in polynomial regression.

Decrease the correlation between X and X^2.