Final Exam

Caleb Ren

Due Tues, Dec. 17 at 11:59pm

Name: **Caleb Ren** Harvard ID: **41291850**

This exam is open-notes and open-book (feel free to use the internet as well). You cannot discuss the contents of the exam with anyone except the teaching staff. There are 5 questions. Please submit your complete exam as a pdf as well as the source R-markdown (or R script) file on Canvas. The exam is **Due at 11:59pm on Tuesday, Dec 17**.

I confirm that I have worked independently on this take-home exam, except for any assistance I may have received from the teaching staff with technical issues. All the work is my own, and I have not collaborated in any way with fellow students.

Signature (include an image of your signature):_____

Problem 1 [25 points total]

Be sure to show your work to receive full credit (or will allow for more partial credit, like in the multiple choice problem).

(a) A 95% confidence interval for the mean, μ, was calculated to be $(-10.2, 0.5)$, and the 90% confidence interval on the same set of data was calculated to be $(-9.4, -0.3)$. Which one of the following would be a possible p-value for the hypotheses: $H_0 : \mu = 0$ vs. $H_A : \mu \neq 0$?

(A) 0.0162

(B) 0.0324

(C) 0.0648

(D) 0.1296

Because the 95% confidence interval for μ contains 0, we would fail to reject the null hypothesis with $\alpha = 0.05$, meaning the p-value is greater than 0.05. Because the 90% confidence interval does not contain 0, we would reject the null hypothesis with $\alpha = 0.10$, meaning the p-value is less than 0.10. Therefore, the only viable option is to have a p-value of 0.03 of 0.0648.

(b) A logistic regression model was fit to predict whether or not someone has ever smoked crack (1 = Yes, 0 = No) from marital status (1 = Married, 2 = Never Married, 3 = anything else: divorced, separated, or widowed) in the General Social Survey (GSS). Relevant R output is shown below:

```
> table(gss$married)
  1   2   3
998 670 678
> round(summary(glm(crack~married,data=gss))$coef,5)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02353    0.00963 2.44378  0.01466
married2     0.06911    0.01496 4.62029  0.00000
married3     0.05605    0.01546 3.62523  0.00030
```

i) Provide a 95% confidence interval for the odds ratio of smoking crack comparing never married individuals to married individuals and briefly interpret this result.

The estimate $e^{\hat{\beta}_1}$ for the odds ratio of smoking crack comparing never married and married is equal to:

$$\widehat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = e^{\hat{\beta}_1}$$

This quantity is equal to 1.0715541. The 95% confidence interval for the slope β_1 is given by $\hat{\beta}_1 \pm 1.96\widehat{SE}_{\beta_1}$. A corresponding 95% confidence interval for the odds ratio e^{β_1} is given by $e^{\hat{\beta}_1 \pm 1.96\widehat{SE}_{\beta_1}}$.

```
b_0 <- 0.02353; b_1 <- 0.06911; b_2 <- 0.05605
se <- 0.01496
```

```r
c(exp(b_1 + qnorm(0.975) * se),
  exp(b_1 - qnorm(0.975) * se))
```

```
## [1] 1.103438 1.040591
```

The 95% confidence interval for $e^{\hat{\beta}_1}$ is given by $\boxed{(1.1034383, 1.0405911)}$.

ii) Determine the observed proportion and number of individuals that smoked crack in each of the 3 marital groups.

The estimate of the intercept in the logistic model $\hat{\beta}_0$ is the log-odds of the proportion when `married2` and `married3` are equal to 0; in other words, when couples are unmarried.

$$\begin{aligned}
\log\left(\frac{\hat{p_0}}{1-\hat{p_0}}\right) &= \hat{\beta}_0 \\
\frac{\hat{p_0}}{1-\hat{p_0}} &= e^{\hat{\beta}_0} \\
\hat{p_0} &= \frac{1}{e^{-\hat{\beta}_0}+1} \\
\hat{p_0} &= 0.5058822
\end{aligned}$$

The estimate of the proportion of individuals who have smoked crack who are married is $\boxed{0.5058822}$. The estimate of the number of people who have smoked crack who are married is $\boxed{505}$.

For the remaining two marriage groups, the odds ratio is the ratio between the odds within the group and the reference group (in this case, the married group).

$$\begin{aligned}
e^{\hat{\beta}} &= \frac{\hat{p_i}/(1-\hat{p_i})}{\hat{p_0}/(1-\hat{p_0})} \\
e^{\hat{\beta}_i} &= \frac{\hat{p_i}/(1-\hat{p_i})}{e^{\hat{\beta}_0}} \\
e^{\hat{\beta}_i+\hat{\beta}_0} &= \frac{\hat{p_i}}{1-\hat{p_i}} \\
\hat{p_i} &= \frac{1}{1+e^{-(\hat{\beta}_i+\hat{\beta}_0)}}
\end{aligned}$$

For those who are unmarried, the proportion of individuals who have smoked crack is $\boxed{0.5231435}$. The number of individuals who have smoked crack who are unmarried is $\boxed{351}$. For those who are other (divorced, separated, or widowed), the proportion of individuals who have smoked crack is $\boxed{0.5198845}$. The number of individuals who have smoked crack who are unmarried is $\boxed{352}$.

(c) Briefly explain what the Normality assumption is in linear regression and briefly explain why this is the assumption that is typically of least concern.

The Normality assumption in linear regression is that the sub-population of responses for each value of the explanatory variable are Normally distributed around the estimated mean conditional on the explanatory variable. In simpler terms, it means that we assume the

errors are Normally distributed. This assumption may have importance when sample size is small, but in large samples, the Normality of error terms has little effect on the overall linear regression. In fact, this is stated in the Gauss-Markov Theorem in which the OLS estimator is the best linear unbiased estimator in terms of minimizing MSE as long as the errors are uncorrelated with mean 0 and constant variance. Notice that there is no statement of Normality in the GM theorem.

(d) An observation is said to have **high leverage** if it is far away from the other observations in the predictor space. Interpret what the effect this has on the standard OLS estimate for $\hat{\vec{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{Y}$. Hint: think about the extreme situation for a simple model.

Under OLS, the slope estimate $\hat{\beta}_1$ is given by $\hat{\vec{\beta}_1} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{Y}$. Therefore, if we were to generate a vector of predicted responses $\hat{\vec{Y}} = \mathbf{X}\hat{\vec{\beta}_1} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{Y} = \mathbf{H}\vec{Y}$, then $\mathbf{H}$ is the mapping from responses $\vec{Y}$ to predicted responses $\hat{\vec{Y}}$.

$$\begin{aligned}
\vec{Y} &= \mathbf{X}\hat{\vec{\beta}_1} + \vec{\epsilon} \\
\vec{Y} - \hat{\vec{Y}} &= \vec{\epsilon} \\
\vec{Y} - \mathbf{H}\vec{Y} &= \vec{\epsilon} \\
(I - \mathbf{H})\vec{Y} &= \vec{\epsilon}
\end{aligned}$$

Let's examine what happens to each regression error ϵ_i. Since we are in an OLS setting, we assume the errors to be uncorrelated with variance σ^2. Given that the matrix $I - \mathbf{H}$ is idempotent ($(I - \mathbf{H})^2 = I - \mathbf{H}$) and symmetric ($(I - \mathbf{H})^T = (I - \mathbf{H})$):

$$\begin{aligned}
\text{Var}(\vec{\epsilon}) &= \text{Var}((I - \mathbf{H})\vec{Y}) \\
&= (I - \mathbf{H})\text{Var}(\vec{Y})(I - \mathbf{H})^T \\
&= \sigma^2(I - \mathbf{H})^T(I - \mathbf{H}) \\
&= \sigma^2(I - \mathbf{H})^2 \\
&= \sigma^2(I - \mathbf{H})\text{Var}(\epsilon_i) \qquad\qquad = \sigma^2(I - \mathbf{H}_{i,i})
\end{aligned}$$

This result shows us that the variance of the ith error is determined and only determined by the magnitude of the ith diagonal element of the projection (or hat) matrix $\mathbf{H}$. Since the term in front of the hat matrix term is negative, as leverage increases (observed point is more of an outlier), we would expect lower variance (and thus less noise). As such, with high leverage observations, we would expect the $\hat{\vec{\beta}_1}$ slope estimate to shrink toward the higher leverage observations, since by minimizing residual error in OLS, we would inadvertantly shrink the slope estimate toward the higher leverage observations. One extreme example would be a dataset that is otherwise clustered around the origin with a single, high-leverage outlier at point $(100, 100)$. An OLS estimate would most likely pass very close to this point in order to minimize residual error, meaning that

(e) Let the sample size be large where the Central Limit Theorem (CLT) typically would hold ($n > 200$). In an ordinary least squares regression model, explain why high leverage observations are a concern when the normality assumption does not hold (be sure to mention what gets affected).

Problem 2 [35 points total]

The dataset `mlb14_18.csv` has team measured statistics for the 30 major league baseball teams from 2014 until 2018, while `mlb19.csv` has the same variables, but for the 2019 season. The variables measured in the training dataset are:

- **team**: a categorical variable with 30 categories to represent the 30 teams. BOS is for the Boston Red Sox while NYY is for the New York Yankees, for example.
- **year**: the year of the measurement, from 2014 until 2018. Note: each team is measured exactly 5 times in the training data set (one measurement for each team for each year).
- **winpct**: the winning percentage for that team in that specific year. By construction, these should be 0.500 on average within each year. This is the response variable.
- **payroll**: the total salary of players who were on the team on the first day of the season that year (measured in millions of dollars).
- **age**: the average age of the players on the first day of the season (weighted by how much they were projected to play).

Note: you are required to treat the 2014-2018 dataset as training data (to fit all models), and the 2019 as testing data (when called for).

(a) Fit an ordinary least squares (OLS) model to predict winpct from team (**lm1**). Formally determine if there is evidence of a difference in winning percentage across teams.

```
training <- read.csv("data/mlb14_18.csv"); testing <- read.csv("data/mlb19.csv")
lm1 <- lm(winpct ~ team, data = training)
summary(lm1)
```

```
##
## Call:
## lm(formula = winpct ~ team, data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18900 -0.03655 -0.00080  0.03775  0.12020
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.47780    0.02771  17.242  < 2e-16 ***
## teamATL     -0.01300    0.03919  -0.332  0.74069
## teamBAL      0.00120    0.03919   0.031  0.97562
## teamBOS      0.06900    0.03919   1.761  0.08085 .
## teamCHC      0.09040    0.03919   2.307  0.02279 *
## teamCHW     -0.03820    0.03919  -0.975  0.33166
## teamCIN     -0.05420    0.03919  -1.383  0.16924
## teamCLE      0.08300    0.03919   2.118  0.03625 *
## teamCOL     -0.00080    0.03919  -0.020  0.98375
## teamDET     -0.00980    0.03919  -0.250  0.80297
## teamHOU      0.07040    0.03919   1.796  0.07496 .
```

```
## teamKCR      0.01960    0.03919   0.500  0.61791
## teamLAA      0.03720    0.03919   0.949  0.34442
## teamLAD      0.10540    0.03919   2.689  0.00818 **
## teamMIA     -0.02380    0.03919  -0.607  0.54481
## teamMIL      0.02160    0.03919   0.551  0.58256
## teamMIN     -0.01500    0.03919  -0.383  0.70259
## teamNYM      0.01980    0.03919   0.505  0.61433
## teamNYY      0.07260    0.03919   1.852  0.06641 .
## teamOAK      0.01240    0.03919   0.316  0.75225
## teamPHI     -0.04200    0.03919  -1.072  0.28601
## teamPIT      0.04300    0.03919   1.097  0.27475
## teamSDP     -0.03840    0.03919  -0.980  0.32914
## teamSEA      0.03560    0.03919   0.908  0.36550
## teamSFG      0.01100    0.03919   0.281  0.77944
## teamSTL      0.07400    0.03919   1.888  0.06141 .
## teamTBR      0.01000    0.03919   0.255  0.79903
## teamTEX      0.00980    0.03919   0.250  0.80297
## teamTOR      0.03320    0.03919   0.847  0.39860
## teamWSN      0.08140    0.03919   2.077  0.03993 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06197 on 120 degrees of freedom
## Multiple R-squared:  0.3717, Adjusted R-squared:  0.2198
## F-statistic: 2.448 on 29 and 120 DF,  p-value: 0.0003809
```

```
anova(lm1)
```

```
## Analysis of Variance Table
##
## Response: winpct
##            Df  Sum Sq   Mean Sq F value    Pr(>F)
## team       29 0.27256 0.0093986  2.4477 0.0003809 ***
## Residuals 120 0.46077 0.0038398
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Categorical regression is equivalent to an ANOVA F-test. With 29 degrees of freedom, we obtain an F-statistic of 2.4477 which corresponds to a p-value of 0.0003809, meaning we reject the null hypothesis. We have evidence to believe that there is a statistically significant difference in winpct between teams.

(b) Fit an OLS model to predict winpct from payroll (**lm2**). Briefly interpret the estimate of β_1 (including statistical significance) and comment on the assumptions of this model.

```
lm2 <- lm(winpct ~ payroll, data = training)
summary(lm2)
```

```
##
## Call:
```

```
## lm(formula = winpct ~ payroll, data = training)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.219890 -0.046946  0.002028  0.039320  0.136606
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.4238166  0.0144325  29.365  < 2e-16 ***
## payroll     0.0006081  0.0001074   5.659 7.65e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06382 on 148 degrees of freedom
## Multiple R-squared:  0.1779, Adjusted R-squared:  0.1723
## F-statistic: 32.02 on 1 and 148 DF,  p-value: 7.645e-08
```

Our null hypothesis is that there is no relationship between payroll and winpct and our alternative hypothesis is that there exists a relationship between payroll and winpct. The OLS estimator β_1 has a value of 0.0006081, a t-value of 5.659, and an associated p-value of 7.65e-08, which is less than our α value of 0.05, meaning we reject the null hypothesis and believe that a statistically significant relationship between payroll and winpct exists. The value of β_1 means that for every \$1 million increase in payroll for a given team, a team can expect on average a 0.06081% increase in winpct. One standard deviation in payroll is around \$50 million, so in more meaningful terms, every \$50 million increase in payroll for a given team is expected to result in about a 3.04% increase in winpct.

(c) Fit an OLS model to predict winpct from payroll and team (**lm3**). Compare the estimated coefficient for payroll's association with the response to its counterpart in **lm2**, and explain what this indicates.

```
lm3 <- lm(winpct ~ payroll + team, data = training)
summary(lm3)
```

```
##
## Call:
## lm(formula = winpct ~ payroll + team, data = training)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.192656 -0.039400 -0.004928  0.039163  0.130967
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4175421  0.0315447  13.237  < 2e-16 ***
## payroll      0.0006354  0.0001807   3.517 0.000619 ***
## teamATL     -0.0191608  0.0374976  -0.511 0.610305
## teamBAL     -0.0248274  0.0381807  -0.650 0.516778
## teamBOS      0.0072657  0.0413658   0.176 0.860871
```

```
## teamCHC      0.0574115  0.0386133   1.487 0.139704
## teamCHW     -0.0397915  0.0374594  -1.062 0.290269
## teamCIN     -0.0530365  0.0374581  -1.416 0.159420
## teamCLE      0.0816281  0.0374587   2.179 0.031291 *
## teamCOL     -0.0072385  0.0375013  -0.193 0.847273
## teamDET     -0.0475462  0.0389640  -1.220 0.224779
## teamHOU      0.0597590  0.0375786   1.590 0.114434
## teamKCR      0.0087870  0.0375826   0.234 0.815539
## teamLAA      0.0024213  0.0387401   0.063 0.950269
## teamLAD      0.0263785  0.0436791   0.604 0.547048
## teamMIA     -0.0123612  0.0375976  -0.329 0.742902
## teamMIL      0.0297254  0.0375278   0.792 0.429884
## teamMIN     -0.0185771  0.0374704  -0.496 0.620964
## teamNYM     -0.0056897  0.0381514  -0.149 0.881699
## teamNYY      0.0046938  0.0421405   0.111 0.911498
## teamOAK      0.0307753  0.0378193   0.814 0.417416
## teamPHI     -0.0509376  0.0375427  -1.357 0.177416
## teamPIT      0.0451250  0.0374615   1.205 0.230760
## teamSDP     -0.0284120  0.0375641  -0.756 0.450928
## teamSEA      0.0078138  0.0382808   0.204 0.838611
## teamSFG     -0.0431111  0.0404936  -1.065 0.289194
## teamSTL      0.0458270  0.0383037   1.196 0.233914
## teamTBR      0.0300902  0.0378897   0.794 0.428688
## teamTEX     -0.0359490  0.0396512  -0.907 0.366434
## teamTOR     -0.0015481  0.0387378  -0.040 0.968189
## teamWSN      0.0360364  0.0396154   0.910 0.364842
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05922 on 119 degrees of freedom
## Multiple R-squared:  0.4308, Adjusted R-squared:  0.2873
## F-statistic: 3.003 on 30 and 119 DF,  p-value: 1.21e-05
```

(d) Fit an OLS model to predict winpct from payroll, team, and the interaction between the two (**lm4**). Formally determine if the interaction term(s) provide significant explanatory power in this model.

```
lm4 <- lm(winpct ~ payroll * team, data = training)
summary(lm4)
```

```
##
## Call:
## lm(formula = winpct ~ payroll * team, data = training)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.179387 -0.028218 -0.002772  0.031716  0.099978
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.750e-01  1.005e-01   3.732 0.000332 ***
## payroll          1.084e-03  1.026e-03   1.056 0.293638
## teamATL         -8.575e-02  1.952e-01  -0.439 0.661456
## teamBAL          3.309e-01  1.902e-01   1.740 0.085364 .
## teamBOS         -2.598e-01  1.828e-01  -1.421 0.158763
## teamCHC          6.332e-02  1.248e-01   0.507 0.613255
## teamCHW         -1.364e-01  1.942e-01  -0.702 0.484258
## teamCIN          5.093e-02  2.046e-01   0.249 0.803937
## teamCLE          9.344e-02  1.316e-01   0.710 0.479434
## teamCOL         -1.998e-01  1.875e-01  -1.066 0.289474
## teamDET         -1.733e-01  1.537e-01  -1.128 0.262395
## teamHOU         -2.927e-03  1.191e-01  -0.025 0.980449
## teamKCR         -5.462e-02  2.028e-01  -0.269 0.788297
## teamLAA          3.147e-01  2.075e-01   1.516 0.132932
## teamLAD          2.463e-01  1.918e-01   1.284 0.202346
## teamMIA          9.944e-02  1.340e-01   0.742 0.459961
## teamMIL         -1.288e-02  1.359e-01  -0.095 0.924683
## teamMIN         -3.912e-01  3.014e-01  -1.298 0.197585
## teamNYM          2.037e-01  1.377e-01   1.480 0.142462
## teamNYY          3.481e-01  1.842e-01   1.889 0.062078 .
## teamOAK         -9.958e-02  1.612e-01  -0.618 0.538390
## teamPHI          4.196e-02  1.302e-01   0.322 0.748099
## teamPIT          4.120e-02  2.442e-01   0.169 0.866419
## teamSDP          4.796e-02  1.246e-01   0.385 0.701270
## teamSEA          1.843e-01  1.681e-01   1.097 0.275785
## teamSFG          4.920e-01  3.756e-01   1.310 0.193508
## teamSTL          3.167e-01  2.994e-01   1.058 0.292980
## teamTBR          1.276e-01  1.522e-01   0.839 0.403928
## teamTEX         -1.134e-01  1.430e-01  -0.793 0.429904
## teamTOR          2.162e-01  2.005e-01   1.078 0.283860
## teamWSN          4.402e-01  2.463e-01   1.787 0.077294 .
## payroll:teamATL  5.954e-04  1.886e-03   0.316 0.753009
## payroll:teamBAL -2.755e-03  1.560e-03  -1.766 0.080834 .
## payroll:teamBOS  1.164e-03  1.292e-03   0.901 0.369976
## payroll:teamCHC -1.990e-04  1.131e-03  -0.176 0.860740
## payroll:teamCHW  9.811e-04  1.976e-03   0.497 0.620681
## payroll:teamCIN -1.109e-03  2.157e-03  -0.514 0.608360
## payroll:teamCLE -1.317e-04  1.324e-03  -0.099 0.920988
## payroll:teamCOL  1.791e-03  1.809e-03   0.990 0.324682
## payroll:teamDET  6.427e-04  1.263e-03   0.509 0.612109
## payroll:teamHOU  4.945e-04  1.154e-03   0.428 0.669379
## payroll:teamKCR  4.986e-04  1.866e-03   0.267 0.789973
## payroll:teamLAA -2.252e-03  1.581e-03  -1.424 0.157770
## payroll:teamLAD -1.258e-03  1.263e-03  -0.996 0.322045
## payroll:teamMIA -1.350e-03  1.510e-03  -0.894 0.373595
## payroll:teamMIL  5.892e-04  1.484e-03   0.397 0.692341
## payroll:teamMIN  3.684e-03  2.998e-03   1.229 0.222413
```

```
## payroll:teamNYM -1.685e-03  1.227e-03  -1.373 0.173046
## payroll:teamNYY -1.940e-03  1.274e-03  -1.522 0.131424
## payroll:teamOAK  2.174e-03  2.138e-03   1.017 0.311799
## payroll:teamPHI -9.109e-04  1.257e-03  -0.725 0.470485
## payroll:teamPIT  5.927e-05  2.627e-03   0.023 0.982047
## payroll:teamSDP -8.762e-04  1.350e-03  -0.649 0.517959
## payroll:teamSEA -1.415e-03  1.402e-03  -1.009 0.315560
## payroll:teamSFG -3.185e-03  2.253e-03  -1.414 0.160871
## payroll:teamSTL -2.089e-03  2.265e-03  -0.923 0.358663
## payroll:teamTBR -1.319e-03  2.041e-03  -0.646 0.520005
## payroll:teamTEX  2.708e-04  1.185e-03   0.229 0.819698
## payroll:teamTOR -1.620e-03  1.540e-03  -1.052 0.295627
## payroll:teamWSN -2.624e-03  1.691e-03  -1.551 0.124338
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05581 on 90 degrees of freedom
## Multiple R-squared:  0.6178, Adjusted R-squared:  0.3672
## F-statistic: 2.465 on 59 and 90 DF,  p-value: 5.466e-05
```

(e) What is the estimated relationship between winpct and payroll in **lm4** for the New York Yankees (NYY)? What is it for the Boston Red Sox (BOS)? Formally determine whether the slopes are significantly different for these two teams.

(f) Fit an OLS model to predict winpct from payroll, team, year, and the interaction between team and year (**lm5**). Which teams are predicted to have the highest and lowest winning percentage in 2019 from this model?

```
lm5 <- lm(winpct ~ payroll + (team * year), data = training)
y_hat <- predict(lm5, testing)
names(y_hat) <- levels(testing$team)
y_hat[which.max(y_hat)]; y_hat[which.min(y_hat)]
```

```
##       NYY
## 0.7132808
```

```
##      BAL
## 0.197697
```

(g) Build a sequential variable selection model (using the combined forward and backward directions using the 'both' argument in R) to predict winpct where the starting model is **lm5**, the lower scope is the intercept-only model and the upper scope is the model with all 2-way interaction (and main effects) of all 4 predictors in the dataset. Evaluate how this model performs in the 2019 data set via MSE.

```
intercepts <- lm(winpct ~ 1, data = training)
interactions <- lm(winpct ~ . * ., data = training)
step_model <- step(lm5,
                   scope = list(lower = formula(intercepts), upper = formula(interactions)),
                   direction = "both", trace = 0)
y_hat_step <- predict(step_model, testing)
```

```
paste("step model MSE:", round(mean((y_hat_step - testing$winpct)^2), 5))
```

```
## [1] "step model MSE: 0.0138"
```

```
paste("lm5 MSE:", round(mean((predict(lm5, testing) - testing$winpct)^2), 5))
```

```
## [1] "lm5 MSE: 0.00596"
```

```
paste("all two-way interactions MSE:", round(mean((predict(interactions, testing) - testir
```

```
## [1] "all two-way interactions MSE: 0.08988"
```

We see that the step model's test MSE is 0.0138, which is higher than that of the linear model using payroll and the interaction between year and team, but lower than the all 2-way effects model. The full interactions model most likely overfit the training data due to the high number of predictors (123!), which is roughly comparable to the number of observations in the dataset. The linear model using payroll, year, and team is more based on reality compared to the stepwise model.

Problem 3 [40 points total]

(a) Fit a linear mixed effects model (LME) with random intercepts to predict winpct where the clusters are defined by the 30 different teams (always assume the teams define the clusters throughout this problem). Call this model **lmer1**. Compare the estimated coefficient for BOS to the appropriate OLS model from problem 2, and briefly explain why this result is the case.

```
require(lme4)
lmer1 <- lmer(winpct ~ 1 + (1 | team), data = training)
summary(lmer1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: winpct ~ 1 + (1 | team)
##    Data: training
##
## REML criterion at convergence: -375
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1884 -0.6509  0.0223  0.7003  2.2485
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  team     (Intercept) 0.001112 0.03334
##  Residual             0.003840 0.06197
## Number of obs: 150, groups:  team, 30
##
```

```
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 0.499980   0.007916   63.16
```

```
predict(lmer1, testing)[testing$team == "BOS"]
```

```
##         4
## 0.5276717
```

```
lm1$coefficients["teamBOS"]
```

```
## teamBOS
##   0.069
```

(b) Fit an appropriate LME model (**lmer2**) to predict winpct from payroll where the overall effect of payroll is important and the average winpct of teams may vary as well as the effect of payroll may vary by team. Hint: you may have to re-center or scale payroll in order to get rid of collinearity with the intercept. Briefly interpret what this model says about the overall average effect of payroll on winpct.

(c) Write out the full model expression for the LME model **lmer2** in mathematical form (as seen in the notes for Lectures 23 and 24); be sure to define your variables. what are the estimates of the parameters in this model?

(d) Formally test whether the effect of payroll is different across the 30 teams in **lmer2**. Hint: you may have to fit another LME model for comparison.

(e) Fit an LME model (**lmer3**) to predict winpct from payroll and year where the intercept is both fixed and random, payroll is treated as only a fixed effect, and year is treated as only a random effect.

(f) Briefly justify the choice for modeling year as only a random effect and not a fixed effect.

(g) Evaluate (via MSE) how **lmer1**, **lmer2**, and **lmer3** (as well as their OLS counterparts) perform on predicting the 2019 data. Present the results in a well-formatted table that has 3 rows (to represent the 3 LME models) and 2 columns (one column for the LME models, and one column for the OLS counterparts). Which of the 6 models performs best at predicting the out-of-sample data? How do you know?

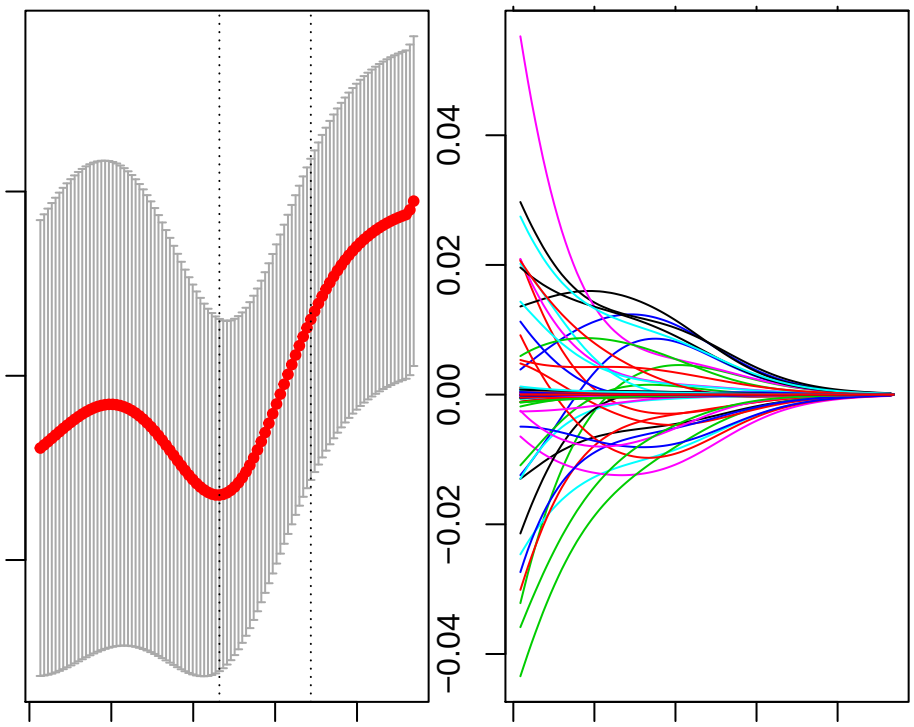
(h) Compare each of the 3 LME models to their OLS counterparts in terms of MSE on test (within each of the 3 pairs, which perform best?) and briefly explain why this is the case for each.

Problem 4 [30 points total]

(a) Build a well-tuned ridge regression model (**ridge1**) to predict winpct from the 4 predictors in the training set along with their 2-way interaction effects. Interpret the relationship between winpct and payroll in this model. Hint: a visual will help.

```
require(glmnet)

X <- model.matrix(interactions, data = training)
ridge1 <- cv.glmnet(X, y = training$winpct, alpha = 0)
par(mfrow = c(1, 2), mar = c(1, 1, 1, 1))
plot(ridge1); plot(ridge1$glmnet.fit, "lambda", label = T)
```

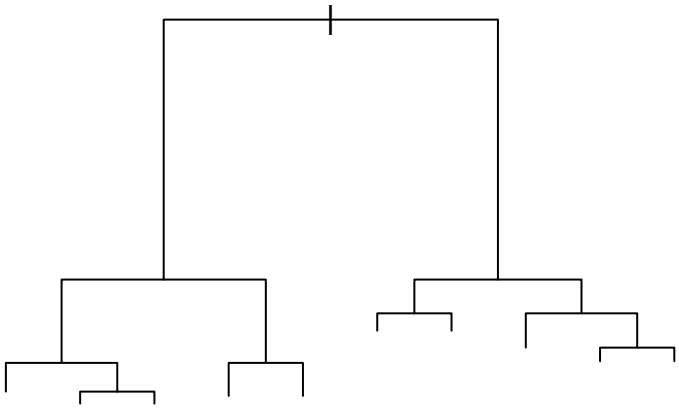



```
ridge <- glmnet(X, y = training$winpct, alpha = 0, lambda = 10^(seq(-4, 4, by = 0.1)))
```

(b) Build a well-pruned decision tree model (**tree1**) to predict winpct from the 4 predictors in the training set. Evaluate which of the predictors are most important within this model.

```
require(rpart)

tree1 <- rpart(winpct ~ ., data = training)
plot(tree1)
```



(c) Build a well-tuned bagging model (**bag1**) to predict winpct from payroll alone. Interpret the relationship between winpct and payroll in this model. Hint: a visual will help.

(d) Build a well-tuned random forest model (**rf1**) to predict winpct from the 4 predictors in the training set. Evaluate which of the predictors are most important within this model.

```
require(randomForest)

rf1 <- randomForest(winpct ~ ., data = training)
```

(e) Which of the predictive models considered for these data is the best (between the 4 predictive models in this problem, the 6 considered in 3(g), and **lm6**)? Briefly justify why this may be the case.

(f) [**Up to 3 points Extra Credit**] Build a prediction model to predict winpct in 2019 that outperforms those considered above. Only use the classes of models we have used in this class (no neural nets or boosted models, for example). The process will be more important than the result. Note: this is not worth very much extra credit.

Problem 5 [20 points total]

In class it was mentioned that a linear regression model can be used to model a binary outcome Y variable, but the logistic regression model is preferred. This simulation steps you through justification why that is the case, from an inferential perspective.

Perform a simulation study (with 2,000 iterations for each of 4 conditions) where the data are generated from the following model:

$$Y_i|X_i \sim \text{Bernoulli}\left(p_i = 0.5 + \delta \cdot (X_i - 0.5)\right)$$

where X_i represents a binary indicator of treatment: with exactly $n/2$ observations in each treatment group. Vary two different parameters (for a total of 4 conditions):

```
i) Use two different treatment effects:  $\delta = 0$ and $\delta = 0.4$.

ii) Use two difference sample sizes: $n = 50$ and $n = 200$.


```r
n <- 50
delta <- 0.4

p5 <- function(n, delta) {
  x <- c(rep(0, n/2), rep(1, n/2))
  p <- c(0.5 + delta * (x - 0.5))
  # n = 50 bernoulli trials
```
```

```
  y <- rbinom(n = n, size = 1, p = p)
  data <- data.frame(x = x, y = y)

  # linear model
  lin_mod <- lm(y ~ x, data)

  # logistic model
  log_mod <- glm(y ~ x, family = "binomial", data = data)
  return(c(summary(lin_mod)$coefficients[2,4],
           summary(log_mod)$coefficients[2,4]))
}
```

(a) In each iteration, analyze the data two different ways: using a linear regression model and separately a logistic regression model, both using the variable X as the sole predictor of Y. Use these to estimate the probability of rejecting the null hypothesis that $H_0 : \beta_1 = 0$ in favor of $H_A : \beta_1 \neq 0$ in each of the two models.

```
p5_p <- list('n:50, d:0' = replicate(2000, p5(50, 0)),
             'n:50, d:0.4' = replicate(2000, p5(50, 0.4)),
             'n:200, d:0' = replicate(2000, p5(200, 0)),
             'n:200, d:0.4' = replicate(2000, p5(200, 0.4)))

p5_lst <- lapply(p5_p,
                 function(x) c(mean(x[1,] < 0.05),
                               mean(x[2,] < 0.05)))
print(p5_lst)
```

```
## $`n:50, d:0`
## [1] 0.0635 0.0450
##
## $`n:50, d:0.4`
## [1] 0.878 0.826
##
## $`n:200, d:0`
## [1] 0.063 0.063
##
## $`n:200, d:0.4`
## [1] 1 1
```

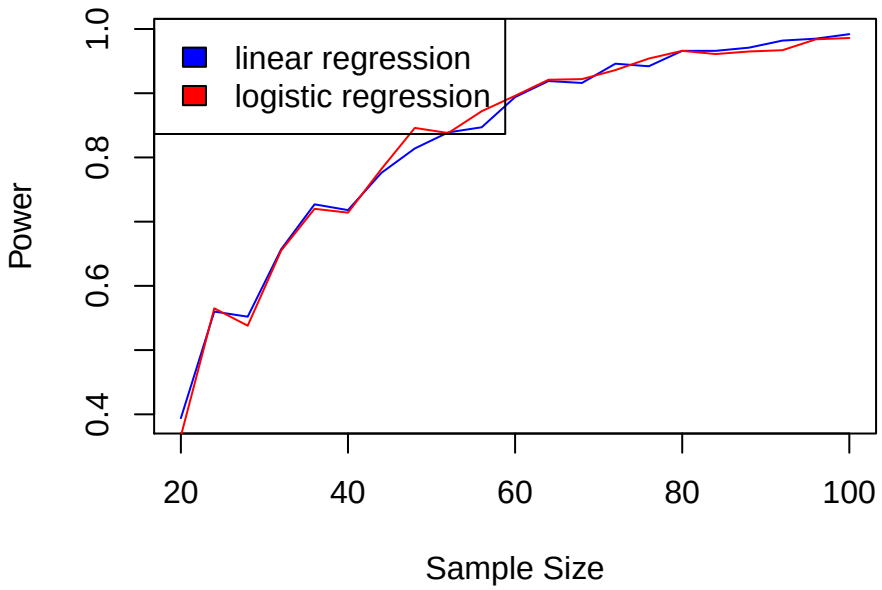
For the $n = 50$ case, we see that when $\delta = 0$, there is low probability (0.0635 and 0.045) that either test is able to pick up differences between both treatments. This is because sample size is comparatively low and there is no true separation in probability of the Bernoulli trials conditional on X ($p = 0.5$ marginally).

For the $n = 50, \delta = 0.4$ case, both linear regression and logistic regression were more powerful (0.878 and 0.826, respectively). Linear regression slightly outperformed logistic regression in this circumstance, but this

(b) Interpret the results: how does sample size affect Type I error rates in the two models? How

does sample size affect statistical power for each?

```
sample_size_lin <- Vectorize(function(n) {
  return(mean(replicate(1000, p5(n, 0.4))[1,] < 0.05))
})
sample_size_log <- Vectorize(function(n) {
  return(mean(replicate(1000, p5(n, 0.4))[2,] < 0.05))
})
ns <- seq(20, 100, by = 4)
prob_lin <- sample_size_lin(ns)
prob_log <- sample_size_log(ns)
plot(ns, prob_lin, col = "blue", type = "l",
     xlab = "Sample Size", ylab = "Power")
lines(ns, prob_log, col = "red")
legend("topleft", legend = c("linear regression", "logistic regression"),
       fill = c("blue", "red"))
```



(c) Explain why Type I error rate is or is not at the nominal 0.05 level for the linear and logistic regression models for the two sample sizes (hint: think about which assumption(s) may be violated).