Final exam, STATS 401 W18

All the questions in this exam refer to the field goal kicking data provided in the R dataframe `goals`. These data record the results of field goal attempts for the kickers who played in all the 2002–2006 National Football League (NFL) seasons. The primary question of interest is whether a kicker who exceeds expectations in one season is likely to do better, or worse, than expected in the following season.

Name. The name of the field goal kicker.

Yeart. The year t corresponding to the row in the dataset.

Teamt. An abbreviation of the name of the team for the kicker in year t.

FGAt. Field goal attempts in year t.

FGt. Percentage of field goal attempts that were successful in year t.

Team.t.1. An abbreviation of the name of the team for the kicker in year $t-1$.

FGAtM1. Field goal attempts in year $t-1$.

FGtM1. Percentage of field goal attempts that were successful in year $t-1$.

Throughout the exam, you may write y_i for the field goal percentage recorded on the ith row of the data file, for $i = 1, \ldots, n$ with $n = 4k$ corresponding to four data points on each of $k = 19$ kickers. You may also write y_{ij} for the jth measurement on kicker i, for $i = 1, \ldots, k$ and $j = 1, \ldots, 4$. You may use this notation without explanation. Other additional notation you use should be defined as appropriate.

```
head(goals)
```

```
##             Name Yeart Teamt FGAt  FGt Team.t.1. FGAtM1 FGtM1
## 1 Adam Vinatieri  2003    NE   34 73.5        NE     30  90.0
## 2 Adam Vinatieri  2004    NE   33 93.9        NE     34  73.5
## 3 Adam Vinatieri  2005    NE   25 80.0        NE     33  93.9
## 4 Adam Vinatieri  2006   IND   19 89.4        NE     25  80.0
## 5    David Akers  2003   PHI   29 82.7       PHI     34  88.2
## 6    David Akers  2004   PHI   32 84.3       PHI     29  82.7
```

---

1. Factors and their coding in R.

We will start the analysis by fitting a basic model, seen earlier in class and homework, specified in R code as

```
lm1 <- lm(FGt~Name+FGtM1, data=goals)
```

(a) *[5 points]*. Write down the sample model fitted by `lm1` in subscript form.

Solution.

$$y_{ij} = m + a_i + b x_{ij} + e_{ij}$$

where x_{ij} is the previous year kicking average for the jth measurement on kicker i, for $i = 1, \ldots, k$ and $j = 1, \ldots, 4$, and e_{ij} is the residual error. m is the estimated intercept, a_i is an additive effect for kicker i, with $a_1 = 0$. b is the estimated coefficient for the effect of the previous year. All coefficients are estimated by least squares.

(b) *[3 points]*. Write down the first 6 rows of the design matrix for `lm1`. You may use dots $(\cdots)$ to abbreviate entries following a repeated pattern, but if you do this it must be clear what they represent.

Solution.

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
## [1,]    1    0    0    0    0    0    0    0    0     0     0     0     0
## [2,]    1    0    0    0    0    0    0    0    0     0     0     0     0
## [3,]    1    0    0    0    0    0    0    0    0     0     0     0     0
## [4,]    1    0    0    0    0    0    0    0    0     0     0     0     0
## [5,]    1    1    0    0    0    0    0    0    0     0     0     0     0
## [6,]    1    1    0    0    0    0    0    0    0     0     0     0     0
##      [,14] [,15] [,16] [,17] [,18] [,19] [,20]
## [1,]     0     0     0     0     0     0  90.0
## [2,]     0     0     0     0     0     0  73.5
## [3,]     0     0     0     0     0     0  93.9
## [4,]     0     0     0     0     0     0  80.0
## [5,]     0     0     0     0     0     0  88.2
## [6,]     0     0     0     0     0     0  82.7
```

```
coef(summary(lm1))["FGtM1",]
```

```
##      Estimate    Std. Error       t value      Pr(>|t|)
## -5.037008e-01  1.127613e-01 -4.466963e+00  3.899977e-05
```

2. Model interpretation. *[4 points]*.

A direct interpretation of the estimated coefficient for the previous year field goal percentage from `lm1` (shown above) is that field goal kickers who kick well one season tend to kick relatively poorly the next season. Explain why general principles for the interpretation of observational studies should make us cautious about jumping to that conclusion.

Solution. All observational studies can be subject to **confounding variables**, which are unmeasured phenomena that affect both the explanatory variables and the response. A particulary type of confounding is **selection bias**. Here, a possible source of selection bias is that these kickers were selected to be those who held their job for 5 consecutive years, so they may not be representative of all kickers. In particular, those who held their job are more likely to have successful earlier seasons, whether by luck or by skill. Another potential confounding variable is coaching strategies. The observational study can readily infer **association** but not so easily establish the **causal mechanism**.

---

3. Model diagnostics.

One possible explanation behind some, or all, of the negative association between kicking percentages in subsequent years could be that coaches who have lower expectation of the abilities of the kicker tend to refrain from hard field goal attempts the following season, pushing up the next season's success rate average. Correspondingly, a coach emboldened by successful kicking may follow this up with choosing to kick in challenging situations. To investigate this, we can consider a linear model where the number of field goal attempts in year t is explained by the field goal success rate in year $t-1$.

```
lm2 <- lm(FGAt~Name+FGtM1, data=goals)
anova(lm2)
```

```
## Analysis of Variance Table
##
## Response: FGAt
```

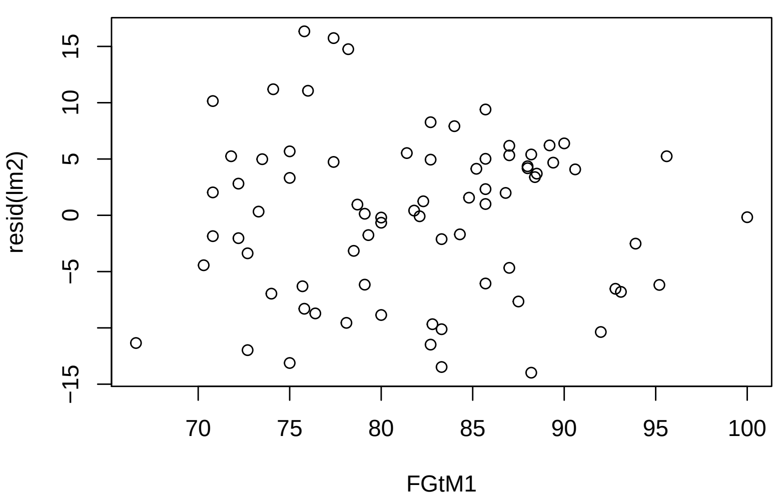
```
##           Df Sum Sq Mean Sq F value Pr(>F)
## Name      18  623.0  34.613  0.5027 0.9459
## FGtM1      1    1.8   1.823  0.0265 0.8713
## Residuals 56 3855.7  68.851
```

(a) *[4 points]*. Interpret the results of this fitted linear model in the context of question of primary interest in the data analysis. You are not asked to give all the details for a hypothesis test or confidence interval. That will come in later questions; here, it is enough to describe briefly the statistical reasoning behind your interpretation.

Solution. According to this ANOVA table, field goal percentage in the previous year does not significantly help in predicting subsequent field goal attempts (F test, p-value = 0.871). Also, kicker does not significantly predict field goal attempts (F test, p-value = 0.946). The identity of the kicker is closely related to the identity of the team, so we also infer that teams do not vary substantially on their field goal kicking strategies. It appears that coaching strategies cannot substantially explain the negative association of primary interest to this investigation.

We should always investigate the data graphically in addition to fitting a model.

```
plot(resid(lm2)~FGtM1, data=goals)
```



(b) *[2 points]*. Comment on your interpretation of the above residual plot, and how it relates to your answer to (a).

Solution. There is no clear nonlinear pattern. The residuals are scattered in a football shape, consistent with the model. This supports the finding from (a), since we can't tell without looking at the data whether there are serious model violations that could make the ANOVA table in (a) unreliable.

One other possibility proposed in class to explain the unexpected results of our first model is that kickers must do well in the earlier years included in the dataset, since they necessarily maintained their position on the team throughout the 2002–2006 interval. The following model investigated the evidence for the magnitude of this effect.

```
lm3 <- lm(FGt~Name+FGtM1+factor(Yeart), data=goals)
anova(lm3)
```

```
## Analysis of Variance Table
##
## Response: FGt
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## Name           18 1569.68   87.20  2.1577   0.01573 *
## FGtM1           1  769.99  769.99 19.0520 5.923e-05 ***
## factor(Yeart)  3   18.97    6.32  0.1564   0.92508
## Residuals      53 2141.99   40.41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. An investigation using an F-test.

(a) *[5 points]*. Write out in full, using subscript form, the alternative hypothesis, H_a, for using `lm3` to test whether the field goal average changes over time.

Solution. Under the alternative hypothesis, the observation y_{ij} is modeled as being generated according to a random variable Y_{ij} constructed according to the equation

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma x_{ij} + \epsilon_{ij}.$$

Here, x_{ij} is the previous year kicking average for the jth measurement on kicker i, for $i = 1, \dots, k$ and $j = 1, \dots, 4$. The measurement error ϵ_{ij} is a normally distributed random variable with mean 0 and standard deviation σ, independent of the other error terms. μ is the intercept, α_i is an additive effect for kicker i, and β_j is an additive effect for year j. We set $\alpha_1 = \beta_1 = 0$ to avoid colinearity. γ is the coefficient for the effect of last year's kicking average.

(b) *[5 points]*. Carry out an F test of the hypothesis H_a against a suitably constructed null hypothesis, H_0, giving explanation of how this test is constructed. What do you conclude?

Solution. The null hypothesis here is $H_0 : \beta_j = 0$ for
The test statistic is

$$f = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_a)/d}{\mathrm{RSS}_a/(n-q)},$$

where d is the difference in degrees of freedom between the null and alternative hypotheses (here, $d = 3$) and q is the degrees of freedom in the alternative hypothesis (here, $q = 1 + 18 + 1 + 3 = 23$), RSS_0 is the residual sum of squares under the null hypothesis, and RSS_a is the residual sum of squares under the alternative. A model-generated version F for this statistic under the null hypothesis has the F distribution on d and $n - q$ degrees of freedom. From the R output, the p-value is $\mathrm{P}(F > f) = 0.925$. The test result is insignificant at the usual 0.05 level. We infer that there is no evidence supporting systematic differences between years in field goal kicking percentage. This is evidence against a role for selection bias in the observed negative estimated value of the coefficient γ.

---

5. A confidence interval.

(a) *[5 points]*. Using the model in Question 1 and the R output on `lm1`, explain how R obtains the estimated coefficient of goal kicking percentage in year $t-1$ as a predictor of goal kicking percentage in year t. Also, using the probability model implicitly assumed in the analysis of Question 1, explain how to the construct a 95% confidence interval for the true coefficient.

Solution. b is computed as a component of $\left(\mathbb{X}\mathbb{X}^{\mathrm{T}}\right)^{-1}\mathbb{X}^{\mathrm{T}}\boldsymbol{y}$. According to the implicit probability model for Question~1, the sample coefficient $b = -0.504$ is an unbiased estimate of a true coefficient β with variance given by the corresponding diagonal term of $\sigma^2\left(\mathbb{X}\mathbb{X}^{\mathsf{T}}\right)^{-1}$ where σ is the standard deviation of the measurement error model and $\mathbb{X}$ is the design matrix. We estimate σ^2 by $s^2 = \frac{1}{n-p}\sum_{i=1}(y_i - \hat{y}_i)^2$ where $\hat{y}_i$ is the fitted value and $p = 20$ is the degrees of freedom. This gives us a standard error which, from the R output provided, is

$$\mathrm{SE} = 0.113.$$

Using a normal approximation, a 95% confidence interval is $[b - 1.96 \times \mathrm{SE}, b + 1.96 \times \mathrm{SE}]$.

(b) *[3 points]*. A confidence interval is only as trustworthy as the model that it is derived from. Explain to what extent you feel the confidence interval is justified based on the analysis available in this exam. Propose any supplementary analysis you would do to strengthen this inference.

The model diagnostics we have seen in questions 2, 3 and 4 all support the linear model of question 1. This gives some support for the resulting confidence interval. We didn't check normality: the scatter of points in the residual plot suggests that there are no outliers and normality is a reasonable approximation. Normality is more critical for prediction than for confidence intervals on coefficients, since in the latter case a cental limit theorem applies: the coefficient estimate is a certain average of all the data points, and averages can be expected to have the *central limit property* that they are well approximated by a normal distribution.

---

6. Collinearity.

Suppose someone suggests that the rest of the team may also be an important component of field goal success. This leads you to try adding to the model a factor for the team in year t with the following consequence.

```r
lm4 <- lm(FGt~Name+Teamt+FGtM1, data=goals)
summary(lm4)
```

```
##
## Call:
## lm(formula = FGt ~ Name + Teamt + FGtM1, data = goals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0807  -3.2025  -0.4982   4.0692  13.2308
##
## Coefficients: (17 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              126.7703    10.6630  11.889  < 2e-16 ***
## NameDavid Akers           -3.6917     4.7822  -0.772   0.4436
## NameJason Elam            -2.0890     4.8118  -0.434   0.6660
## NameJason Hanson           3.1180     4.7613   0.655   0.5154
## NameJay Feely             -5.2243     5.7213  -0.913   0.3654
## NameJeff Reed             -7.3385     4.7801  -1.535   0.1308
## NameJeff Wilkins           3.2869     4.7674   0.689   0.4936
## NameJohn Carney           -5.0437     4.8041  -1.050   0.2986
## NameJohn Hall             -7.5838     4.8506  -1.563   0.1240
## NameKris Brown           -12.4942     4.9275  -2.536   0.0143 *
## NameMatt Stover            9.7595     4.7649   2.048   0.0456 *
## NameMike Vanderjagt        3.6936     7.2192   0.512   0.6111
## NameNeil Rackers          -5.6610     4.7785  -1.185   0.2415
## NameOlindo Mare          -12.1338     4.8506  -2.501   0.0156 *
## NamePhil Dawson            4.5452     4.7621   0.954   0.3443
## NameRian Lindell          -3.9423     4.8153  -0.819   0.4167
## NameRyan Longwell         -5.2597     7.3294  -0.718   0.4762
```

```
## NameSebastian Janikowski  -3.0388     4.7995  -0.633   0.5294
## NameShayne Graham          3.1111     4.7677   0.653   0.5169
## TeamtATL                  -8.4916     6.2682  -1.355   0.1814
## TeamtBAL                       NA         NA      NA       NA
## TeamtBUF                       NA         NA      NA       NA
## TeamtCIN                       NA         NA      NA       NA
## TeamtCLE                       NA         NA      NA       NA
## TeamtDAL                  -2.9588    10.1814  -0.291   0.7725
## TeamtDEN                       NA         NA      NA       NA
## TeamtDET                       NA         NA      NA       NA
## TeamtGB                    5.3209     7.3222   0.727   0.4707
## TeamtHOU                       NA         NA      NA       NA
## TeamtIND                   3.9384     7.2302   0.545   0.5883
## TeamtMIA                       NA         NA      NA       NA
## TeamtMIN                       NA         NA      NA       NA
## TeamtNE                        NA         NA      NA       NA
## TeamtNO                        NA         NA      NA       NA
## TeamtNYG                       NA         NA      NA       NA
## TeamtOAK                       NA         NA      NA       NA
## TeamtPHI                       NA         NA      NA       NA
## TeamtPIT                       NA         NA      NA       NA
## TeamtSTL                       NA         NA      NA       NA
## TeamtWAS                       NA         NA      NA       NA
## FGtM1                     -0.5164     0.1170  -4.414 5.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.234 on 52 degrees of freedom
## Multiple R-squared:  0.551,  Adjusted R-squared:  0.3524
## F-statistic: 2.774 on 23 and 52 DF,  p-value: 0.00117
```

(a) *[4 points]*. Explain why all but four of the coefficients for the team factors take value `NA`.

Solution. `NA` estimates occur when the columns of the design matrix $\mathbb{X}$ are collinear. Here, we can expect considerable collinearity between the team and the kicker. If a kicker stays with the same team throughout the 4 years of the dataset, the effect of this team and this kicker are indistinguishable. This is equivalent to the corresponding columns of the design matrix being collinear. The effects only become distinguishable when the kicker changes team, and this apparently occurs on only 4 occasions.

The following results show that if we put the kicker into the model first, then the team appears insignificant from an F test. However, if we put team first then it is significant and kicker becomes insignificant.

```r
anova(lm(FGt~Name+Teamt+FGtM1, data=goals))
```

```
## Analysis of Variance Table
##
## Response: FGt
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Name      18 1569.68   87.20  2.2440    0.0121 *
## Teamt      4  153.02   38.25  0.9844    0.4242
## FGtM1      1  757.14  757.14 19.4831 5.147e-05 ***
## Residuals 52 2020.79   38.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm(FGt~Teamt+Name+FGtM1, data=goals))
```

```
## Analysis of Variance Table
##
## Response: FGt
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Teamt     21 1721.49   81.98  2.1094   0.01508 *
## Name       1    1.20    1.20  0.0310   0.86100
## FGtM1      1  757.14  757.14 19.4831 5.147e-05 ***
## Residuals 52 2020.79   38.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b) *[4 points]*. Explain why the significance of the effect of the team and the kicker depends on the order in which the variables occur in the model. Can the data distinguish whether the goal kicking percentage is best explained by team or by kicker or by both?

Solution. As discussed in 6(a), the team and the kicker are almost indistinguishable—they are perfectly indistinguishable unless the kicker changes team. If we include the `Name` variable first, and then test for the inclusion of `Teamt` we are testing whether the team adds significantly more explanatory power over the kicker identity, which it doesn't (since both carry almost the same information). However, if we include `Teamt` first, it shows up as significant, as expected since we have already found that `Name` alone is significificant.

The data alone do not establish whether the observed effect is due to the kicker or due to some other aspect of the team, since `Teamt` alone explains the data similary well to `Name` alone. Some common knowledge about football may suggest that the kicker is more responsible for the success of the goal attempt than his team mates, but that conclusion doesn't come directly out of these data.

Acknowledgments: The `goals` data were presented by *A Modern Approach to Regression with R* by S. J. Sheather, and originally come from http://www.rorotimes.com/nfl/stats.
