Lab 11: Multiple Linear Regression

Lab Goals

The goals of this lab are to:

1) Analyze data using MLR
2) Perform some basic inference for regression coefficients
3) Look at some basic ways to compare models

Multiple Linear Regression

A multiple linear regression model is used when more than one predictor variable is assumed to explain the variability in a response, Y. If the assumptions for a MLR model are satisfied, the relationship between each observation, Y_i and the predictors, $X_{i1}, ..., X_{ik}$ is specified as

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_i$

where

1. β_0 is the intercept of the regression
2. β_j is the regression coefficient (or partial slope) associated with the j-th predictor
3. ϵ_i is the error term for observation Y_i
4. $\epsilon_i \sim N(0, \sigma_\epsilon)$ for all $i = 1, \ldots, n$ and are mutually independent

Speed Dating Data

Between 2002 and 2004 a series of speed dating experiments were conducted at Columbia University. Participants were students at Columbia's graduate and professional schools. Each participant attended one speed dating session in which they met with each participant of the opposite sex for four minutes. The data for this analysis contains the responses from 233 speed dates for either the male or the female on how they rated their date from 0 to 10 on the following attributes: *attractive*, *sincere*, *intelligent*, *fun*, *ambitious*, and *shared interests*. Here we will consider the ratings for each of these attributes along wih the *gender* of the participant giving the responses as predictors for a MLR model where the response (Y) is a rating on a scale from 0 to 10 of how much the participant liked their date. The following table includes a description of each predictor.

Predictor	Description
`gender`	M = Male F = Female
`attractive`	attractiveness on a scale of 0-10
`sincere`	sincerity on a scale of 0-10
`intelligent`	intelligence on a scale of 0-10
`fun`	how fun on a scale of 0-10
`ambitious`	how ambitious on a scale of 0-10
`sharedinterests`	interests shared on a scale of 0-10

Read the data into this lab document and use the `names()` function to determine the variable names of this dataset.

```
SpeedDating2 <- read.csv("SpeedDating2.csv")
names(SpeedDating2)
```

```
## [1] "gender"          "like"            "attractive"      "sincere"
## [5] "intelligent"     "fun"             "ambitious"       "sharedinterests"
```

Problem 1

Initially we will consider a model that includes all of the predictors in the table above except for `intelligent` and `ambitious`.

Write out the model. Include definitions of each predictor and regression coefficient. Assume `F` is the first level of `gender`. In R, the covariate for a binary predictor is 0 if the observation belongs to the first level of the predictor and 1 if the observation belongs to the second level of the predictor.

With 5 covariates, the model is,

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i$

where

Y_i = how much the participant i likes their date on a scale of 0 to 10

X_{i1} is equal to 1 if the ith participant is male and 0 if it is female

X_{i2} = attractiveness of the ith participant's date on a scale of 0 to 10

X_{i3} = sincerity of the ith participant's date on a scale of 0 to 10

X_{i4} = how fun was the ith participant's date on a scale of 0 to 10

X_{i5} = interests shared with the ith participant's date on a scale of 0 to 10

β_0 is the expected value of `like` for females when all continuous predictors are 0

β_1 is the mean effect on `like` due to `gender = M`

β_2 is the mean effect on `like` of a one unit increase in `attractive` given all other predictors are set at the same level

β_3 is the mean effect on `like` of a one unit increase in `sincere` given all other predictors are set at the same level

β_4 is the mean effect on `like` of a one unit increase in `fun` given all other predictors are set at the same level

β_5 is the mean effect on `like` of a one unit increase in `sharedinterests` given all other predictors are set at the same level

ϵ_i is the error term associated with observation i

Problem 2

Here we will consider the model defined in Problem 1.

a) Before we fit the linear model in R, we want to make sure that `F` is defined as the first level of the categorical predictor, `gender`. If `M` is the first level of `gender`, as in the past, we can use the `factor()` function to re-order the levels. The following code will list the levels of `gender`; change the ordering of the levels if it is necessary.

```
levels(SpeedDating2$gender)
```

```
## [1] "F" "M"
```

b) Using `like` as the response, the following code will fit the linear model using the `lm()` function and include a summary of the fit.

```
like.lm = lm(like~gender+attractive+sincere+fun+sharedinterests, data=SpeedDating2)
summary(like.lm)
```

```
##
## Call:
## lm(formula = like ~ gender + attractive + sincere + fun + sharedinterests,
##     data = SpeedDating2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3259 -0.5118  0.1204  0.5895  4.8121
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.29149    0.39046   0.747  0.45613
## genderM         -0.05960    0.14640  -0.407  0.68429
## attractive       0.41377    0.04943   8.370 5.92e-15 ***
## sincere          0.15115    0.04990   3.029  0.00274 **
## fun              0.13831    0.04993   2.770  0.00606 **
## sharedinterests  0.24694    0.04053   6.093 4.71e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.107 on 227 degrees of freedom
## Multiple R-squared:  0.6565, Adjusted R-squared:  0.649
## F-statistic: 86.78 on 5 and 227 DF,  p-value: < 2.2e-16
```

c) Using this model, what is the equation that estimates the expected value of `like` given the 5 predictors?

$\hat{E}(Y|predictors) = 0.2915 - 0.0596$*genderM*$+ 0.4138$*attractive*$ + 0.1511$*sincere*$ + 0.1383$*fun*$ + 0.2469$*sharedinterests*

d) What is the estimated expected value of `like` when `gender=F` and all other predictors are zero?

This is the estimated intercept term equal to 0.2915.

e) What is the value of X_{i1} when `gender=F`? What is the value of X_{i1} when `gender=M`?

$X_{i1} = 0$ when `gender=F`. $X_{i1} = 1$ when `gender=M`.

f) What is the estimated expected value of `like` when `gender=M` and all other predictors are zero?

This is the intercept plus the coefficient for *genderM*, *0.2915 - 0.0596 = 0.2319.*

g) What is the estimated expected value of `like` for a female respondant that rated the male a 7 for attractiveness, a 5 for sincerity, an 8 for fun, and a 0 for shared interests?

$\hat{E}(Y) = 0.2915 - 0.0596(0) + .4138(7) + .1511(5) + 0.1383(8) + 0.2469(0) = 5.05$

h) What is the R^2 value for this model? Interpret it in the context of this study. Would we expect the estimate in (g) to be very accurate based on this R^2 value?

$R^2 = 0.657$. *This means about 65% of the variability in* *like* *can be explained by the model. Since a large proportion of variability in* *like* *cannot be explained by the model, we would not expect that the estimate in (g) to be very accurate.*

i) A more specific way to characterize the variability in $E(Y)$ from part (g) is to use the `predict()` function. In the following code chunk, the `predict()` function is first used to create a confidence

interval for $E(Y)$ given the predictors from part (g). Then, the `predict()` function is used to create a prediction interval for a new observation of `like` when the predictors are as specified in part (g).

```
predict(like.lm, list(gender = "F", attractive = 7, sincere = 5, fun = 8, sharedinterests = 0),
    interval = "conf")
```

```
##        fit      lwr      upr
## 1 5.050108 4.430672 5.669544
```

```
predict(like.lm, list(gender = "F", attractive = 7, sincere = 5, fun = 8, sharedinterests = 0),
    interval = "pred")
```

```
##        fit      lwr      upr
## 1 5.050108 2.782246 7.317971
```

Hypothesis Tests Found in the Summary Table

The `summary()` function not only provides estimates for all regression coefficients in the model, $\beta_j, j = 0, ..., k$, but also tests $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$ for $j = 0, .., k$ assuming all other covariates listed are included in the model.

For each of these hypothesis tests, the standardized test statistic is $\frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)}$. Under the null hypothesis, this statistic has a t distribution with $n - p$ degrees of freedom where

1. n = total number of observations, and
2. $p = k + 1$ = total number of regression coefficients in the model including the intercept.

The summary table lists the realization of this statistic under `t-value`. The p-value for the test is $2P(t_{n-p} > |\frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)}|)$ and is listed under `Pr(>|t|)`.

Problem 4

Assuming all predictors but `gender` are set at the same value, is there a significant difference in the expected value of `like` for males versus females?

No, the p-value for the effect of being male (in comparison to being female) is 0.6843. Thus, there is not enough statistical evidence to conclude that the expected value of ***`like`*** *assuming all other predictors are the same is different for females and males.*

However, there still could be differences due to gender in this model. For instance, what a male respondant means by "fun" could be different from what a female means by "fun".

Confidence Intervals

100(1-α)% confidence intervals for each regression coefficient, $\beta_j, j = 0, ..., k$, can be constructed in the typical way:

100(1-α)% confidence interval for $\beta_j = \hat{\beta}_j \pm t_{\alpha/2,n-p}\hat{se}(\hat{\beta}_j)$

Problem 5

Construct a 95% confidence interval for the partial slope of `sharedinterests` and interpret it in the context of the study. Use the `confint()` function to do this:

```
confint(like.lm)
```

```
##                      2.5 %    97.5 %
## (Intercept)     -0.47790481 1.0608779
## genderM         -0.34807142 0.2288649
## attractive       0.31636306 0.5111766
## sincere          0.05282331 0.2494711
## fun              0.03993596 0.2366882
## sharedinterests  0.16707263 0.3268010
```

We are 95% confident that the partial slope for shared interests is between 0.17 and 0.33.

F-test for Comparing a Complete to a Reduced Model

Suppose a MLR model is used to predict the response variable (Y) using k predictors. We might want to compare this model to a *nested model* that includes only a subset of g of the original k predictors. So given the complete model

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_i$.

Suppose we want to test whether the predictors $X_{ig+1}, .., X_{ik}$ are significantly associated with the response variable (Y) given $X_{i1}, .., X_{ig}$ are included in the model. The null hypothesis for this test is

$H_0 : \beta_{g+1} = ... = \beta_k = 0$

The alternative hypothesis for this test is that at least one of these regression coefficients is not equal to zero.

Under the null hypothesis, $F^* = \frac{(SSE_{reduced} - SSE_{complete})/(k-g)}{MSE_{complete}} \sim F_{k-g,n-p}$ where

1. $SSE_{reduced}$ is the error sum of squares for the reduced model
2. $SSE_{complete}$ is the error sum of squares for the complete model
3. k = model degrees of freedom for the complete model
4. g = model degrees of freedom for the reduced model
5. $MSE_{complete}$ is the MSE for the complete model
6. n = total number of observations
7. p = number of regression coefficients in the complete model

For this test, H_0 is rejected if $F^* > F_{\alpha,k-g,n-p}$ (which is the $1-\alpha$ quantile of an F-distribution with $k-g$ and $n-p$ degrees of freedom).

Problem 6

Here we will consider adding the predictors, `intelligent` and `ambition`, to the MLR model for `like`. The code below will fit the complete linear model in R.

```
complete.lm = lm(like ~ gender + attractive + sincere + fun + sharedinterests +
    intelligent + ambitious, data = SpeedDating2)
```

a) The `anova()` function will perform an F test to compare the reduced and complete models. Run the following code to perform this test.

```
anova(like.lm,complete.lm,test="F")
```

```
## Analysis of Variance Table
##
## Model 1: like ~ gender + attractive + sincere + fun + sharedinterests
## Model 2: like ~ gender + attractive + sincere + fun + sharedinterests +
##     intelligent + ambitious
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    227 278.26
## 2    225 273.83  2    4.4259 1.8183 0.1647
```

b) What is the null hypothesis of the test run in (a)?

H_0: the regression coefficients associated with ***intelligent*** *and* ***ambitious*** *are both equal to zero.*

c) What is the p-value of this test? At a 0.05 significance level, should we reject the null hypothesis?

The p-value of the test is 0.16. Since $0.16 > 0.05$*, we will fail to reject the null hypothesis that the regression coefficients associated with* ***intelligent*** *and* ***ambitious*** *are both equal to zero.*

d) Based on (c) which is the preferred model (complete or reduced)?

We would prefer the reduced model since there is not enough statistical evidence that the partial slopes associated with ***intelligence*** *and* ***ambitious*** *are different from zero.*

Problem 7

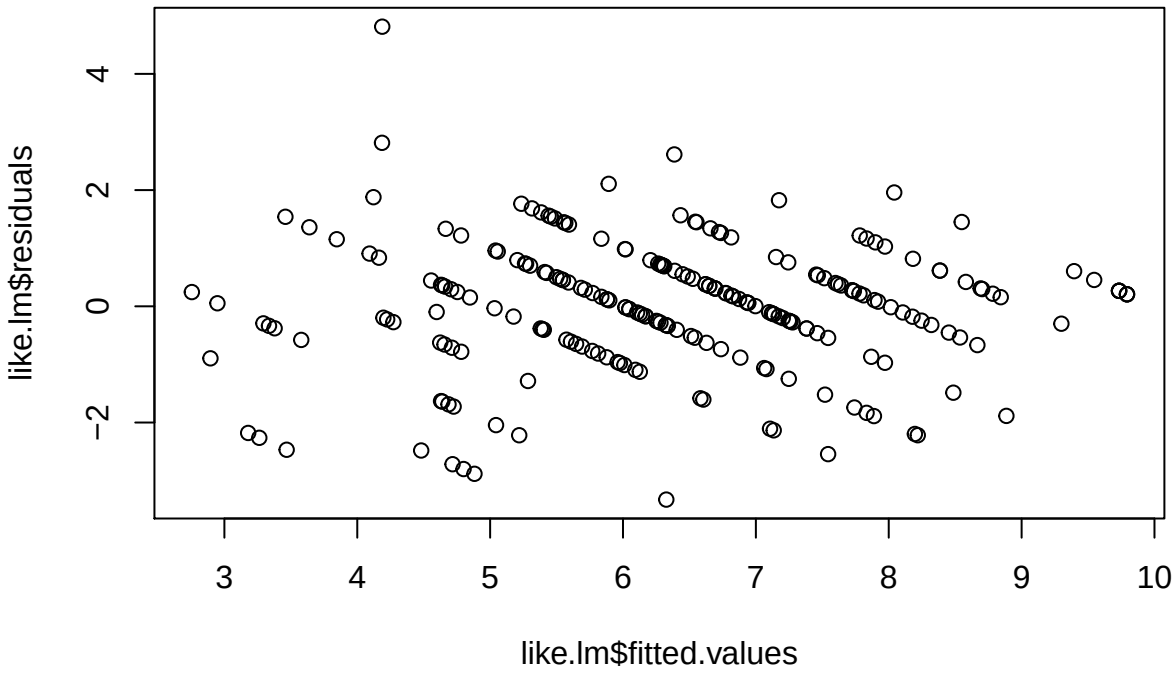
Here we will do a quick check of the assumptions for this MLR model.

a) Does it seem reasonable to assume these observations are independent?

It would be reasonable to assume that every date would be independent of every other date if these participants are randomly sampled (i.e., they weren't friends who were communicating with each other) and each participant were involved in only a single date. However, if each participant was involved in multiple dates that were included in the dataset, then this could lead to dependence in the errors (e.g., the model might be wrong in the same way for all the dates involving that participant.

b) Create a scatterplot of the residuals by the fitted values. Does the equal variance assumption seem reasonable?

```
plot(like.lm$fitted.values, like.lm$residuals)
```

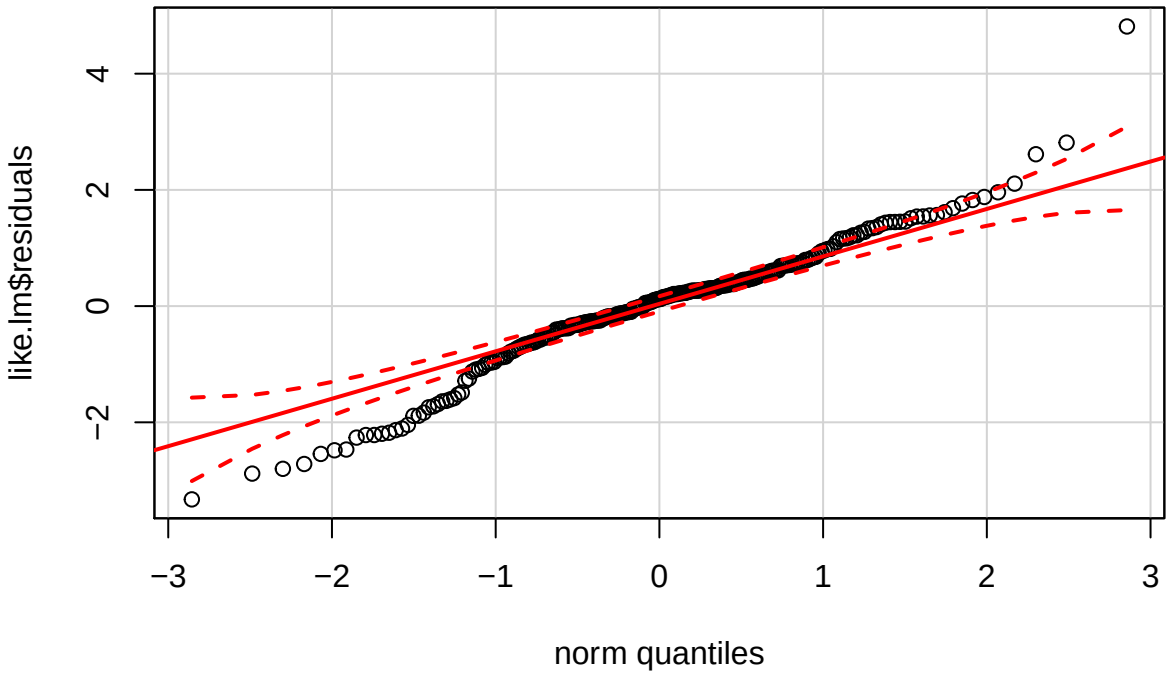


While there are a couple of outliers, the residuals are fairly equally spread around zero. The equal variance assumption is probably reasonable.

c) Create a Q-Q plot of the residuals. Does the normality assumption seems reasonable?

Many points fall outside of the confidence intervals of the Q-Q plot. The normality assumption does not seem reasonable.

```
library(car)
qqPlot(like.lm$residuals)
```



d) Based on your answers to (b) and (c), does it make sense to try and transform either the predictors or the response (or both)? If so, what data transformation would you suggest?

While the normality assumption is not valid, sample size is fairly large and it may be reasonable to interpret the results of this model. However, a data transformation might help correct the violation of the normality assumption. Since some of the variables had 0 responses, a square root transformation may be appropriate here.

Note: Remember that the residuals, $r_i = y_i - \hat{\mu}(x_i)$, are not the same thing as the $\epsilon_i = y_i - \mu(x_i)$. Thus, the residuals not looking normal does not necessarily mean that the ϵ_i are not normal (which in itself would not be so bad since the CLT would help if n is large). When might we expect r_i to be very different from ϵ_i? This can occur when our model is mis-specified. For example, maybe the expected value of Y depends on a predictor we didn't measure. Or the dependence on one or more of the predictors is not linear.

Regression is still a useful tool for making predictions even if assumptions are violated. What becomes invalid is the inferences (e.g., t-tests, F-tests, p-values, confidence intervals) we draw, which were based on our assumptions. If, in truth, the model was properly specified in all ways except for normality of the ϵ_i (i.e., we aren't missing predictors, the dependence is linear, and the ϵ_i are independent), then the CLT would still allow us to make the inferences (as long as n is large enough and ϵ_i's distribution is not too skewed).

However, it is important to remember that when we observe the residuals are not normal, this can indicate more serious problems such as missing a predictor.

More advanced topic (for those interested to try on their own): Interactions

Based on our first lab in which we visualized this data set, we have reason to think that men and women should have a different partial slope for `attractive`. This is referred to as there being an "interaction" between the variables `gender` and `attractive`. Here is an example in which we add two interactions. Observe that both are highly significant:

```
like.lm2 = lm(like ~ . + gender:attractive + gender:fun, data=SpeedDating2)
summary(like.lm2)
```

```
##
## Call:
## lm(formula = like ~ . + gender:attractive + gender:fun, data = SpeedDating2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1490 -0.5038  0.0208  0.6217  4.3414
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.52788    0.52094   1.013 0.312002
## genderM            -1.08874    0.56907  -1.913 0.057003 .
## attractive          0.26566    0.05797   4.583 7.64e-06 ***
## sincere             0.08271    0.05554   1.489 0.137840
## intelligent         0.14304    0.06947   2.059 0.040673 *
## fun                 0.23311    0.06204   3.757 0.000219 ***
## ambitious          -0.06973    0.05175  -1.348 0.179161
## sharedinterests     0.24557    0.03903   6.292 1.64e-09 ***
## genderM:attractive  0.41504    0.09559   4.342 2.14e-05 ***
## genderM:fun        -0.25709    0.08713  -2.951 0.003508 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.064 on 223 degrees of freedom
## Multiple R-squared:  0.6886, Adjusted R-squared:  0.6761
```

```
## F-statistic:  54.8 on 9 and 223 DF,  p-value: < 2.2e-16
```

In BTRY6020/ILRST6200 you'll learn more about interaction models and how to interpret them. From this output, we find that a unit increase in `attractive` has a statistically significantly different effect on `like` for men versus women. In particular, we estimate that the partial slope on `attractiveness` is larger by 0.415 for men than for women. There is also a difference in the slope of `fun` between men and women. We estimate the partial slope on `fun` to be 0.257 higher for women than for men.