BIOS7345 Lab 2

Least-squares regression

Sarah Lotspeich

September 14, 2018

Prologue: Pipes

Before we begin, there are two basic pipes functions I will be using.

- Single pipe: `%>%`
- Double pipe: `%<>%`

The `%>%` operator "pipes" an argument into a function. For example, the following two lines of code are equivalent.

```
# without pipes
mean(seq(1, 10))
```

```
## [1] 5.5
```

```
# with pipes
seq(1, 10) %>% mean()
```

```
## [1] 5.5
```

The single pipe is super useful when you have lots of nested arguments, and it saves you a lot of confusing parentheses.

The `%<>%` operator "pipes" the object on the left-hand side into a function, and then reassigns the new value to the original name. Some simple code:

```
# create a vector
x <- seq(1, 10)

# say I want to squareroot transform the whole vector without pipes
x <- sqrt(x)
x
```

```
##  [1] 1.000000 1.414214 1.732051 2.000000 2.236068 2.449490 2.645751
##  [8] 2.828427 3.000000 3.162278
```

```
# create a vector
x <- seq(1, 10)
# with pipes
x %<>% sqrt()
x
```

```
##  [1] 1.000000 1.414214 1.732051 2.000000 2.236068 2.449490 2.645751
##  [8] 2.828427 3.000000 3.162278
```

Data: Entertainer Career Arcs

These data come from a super cool social media site for data people called data.world. The link to the documentation can be found here, but the TLDR is:

1. we have a sample of 63 notable entertainers (actors, actresses, musicians, etc.) from the last hundred years, and
2. for each we know gender (`gender_traditional`), the age when they first "broke through" (`breakthrough_age`), and the age when they won their first award (`first_award_age`).

Spoiler alert: you won't know all of the names. Darn millenials.

```
stars <- read.csv(file = "https://query.data.world/s/xsweyho62srvyorzoly3w2uctn76rz",
    header = TRUE, stringsAsFactors = FALSE)
```

Objective

We are interested in the following research questions:

1. What is the association between the age at which an entertainer broke through and the age at which they won their first award?
2. Do men and women have different expected ages at which they would win their first award?
3. Is the association between ages of breakthrough and first award different for men and women?

We can answer them using three separate, pre-specified regression models.

Models

Simple linear regression with one continuous predictor

The first model we will fit is for `first_award_age` on `breakthrough_age`. Specify the model, using proper LaTex, below:

$$E(\text{first award age}|\text{breakthrough age}) = \beta_0 + \beta_1(\text{breakthrough age}).$$

Now, begin fitting the model by restructuring these data into your response vector (`y`) and design matrix (`X`). I recommend using `data.matrix()` for this.

```
# create response vector
y <- matrix(data = stars$first_award_age, ncol = 1)

# create design matrix using data.matrix()
X1 <- stars %>% dplyr::mutate(Int = 1) %>% dplyr::select(Int, breakthrough_age) %>%
    data.matrix()
```

Begin by estimating the model coefficients, $\hat{\boldsymbol{\beta}} = \begin{bmatrix}\hat{\beta}_0 & \hat{\beta}_1\end{bmatrix}^T$ according to the Rencher Theorem 7.3a:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

```
# use matrix operations to calculate least-squares estimates beta0, beta1
beta_hat1 <- solve(t(X1) %*% X1) %*% t(X1) %*% y
# for later, go ahead and round these to 3 digits
beta_hat1 %<>% round(3)
```

Now we need an unbiased estimate for the variance of the error terms (σ^2), which we obtain using Rencher Equation (7.23):

$$s^2 = \frac{1}{n-k-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

where $k = $ the number of predictors in your model and $n = $ sample size.

```r
n <- nrow(X1)  #sample size
k <- ncol(X1) - 1  #number of predictors (excl. intercept)
# use (7.23) to estimate the variance of the error terms
s2 <- (1/(n - k - 1)) * t(y - X1 %*% beta_hat1) %*% (y - X1 %*% beta_hat1)
```

The last thing this model needs are measures of uncertainty (e.g. standard errors) for the coefficient estimates. Using our estimate s^2 above, we can calculate these directly according to Rencher (7.27):

$$\hat{Cov}(\hat{\boldsymbol{\beta}}) = s^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

```r
# estimate the SEs for the estimated coefficients
cov_beta_hat <- as.numeric(s2) * solve(t(X1) %*% X1)
# round these off
cov_beta_hat %<>% round(3)
```

To wrap up this model let's make a nice table and print it using `xtable()` (which we totally remember from last week's lab). Recall: what do we need to add in the chunk header to make this print nicely?

```r
# create a table for the model
tab <- data.frame(Variable = c("Intercept", "Breakthrough age"), Effect = beta_hat1,
    SE = round(diag(sqrt(cov_beta_hat)), 3))
```

Try using the `dplyr::mutate()` function to append columns for the Wald Z test statistic and p-value for significance tests for $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 1$. Consider using `ifelse()` to clean up the p-values if they're really small.

```r
# append Z test statistics and p-values
tab %<>% mutate(Z = round(Effect/SE, 3), P = ifelse(pnorm(abs(Z), lower.tail = FALSE) <
    0.001, "<0.001", round(pnorm(abs(Z), lower.tail = FALSE), 3)))
# print model table
tab %>% xtable(caption = "Modeling age at first award on age at breakthrough.")
```

	Variable	Effect	SE	Z	P
1	Intercept	9.36	6.54	1.43	0.076
2	Breakthrough age	1.06	0.22	4.80	<0.001

Table 1: Modeling age at first award on age at breakthrough.

Do future you a favor and put this into a function called `my_ls()` with parameters for the response vectors, `y`, the design matrix, `X`, and a caption for the model table to be printed.

```r
my_ls <- function(y, X, cap) {
    # coefficients ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ use
    # matrix operations to calculate least-squares estimates beta0, beta1
    beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y %>% round(3)
    # variance of the errors ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
    n <- nrow(X)  #sample size
    k <- ncol(X) - 1  #number of predictors (excl. intercept)
    # use (7.23) to estimate the variance of the error terms
    s2 <- (1/(n - k - 1)) * t(y - X %*% beta_hat) %*% (y - X %*% beta_hat)
    # covariance matrix of beta0, beta1 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
    cov_beta_hat <- as.numeric(s2) * solve(t(X) %*% X) %>% round(3)
    # create the table ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ create a
    # table for the model
    tab <- data.frame(Variable = c("Intercept", "Gender"), Effect = beta_hat,
        SE = round(diag(sqrt(cov_beta_hat)), 3))
```

```
    # append Z test statistics and p-values
    tab %<>% dplyr::mutate(Z = round(Effect/SE, 3), P = ifelse(pnorm(abs(Z),
        lower.tail = FALSE) < 0.001, "<0.001", round(pnorm(abs(Z), lower.tail = FALSE),
        3)))
    # print model table
    return(xtable(tab, caption = cap))
}
```

For comparison, you may now fit this model using `rms::ols()`.

```
ols(formula = first_award_age ~ breakthrough_age, data = stars)
```

```
## Linear Regression Model
##
##  ols(formula = first_award_age ~ breakthrough_age, data = stars)
##
##                 Model Likelihood     Discrimination
##                    Ratio Test           Indexes
##  Obs       63    LR chi2     20.08    R2       0.273
##  sigma12.4646    d.f.            1    R2 adj   0.261
##  d.f.      61    Pr(> chi2) 0.0000    g        8.332
##
##  Residuals
##
##      Min      1Q  Median      3Q     Max
##  -12.732  -8.888  -5.670   5.487  38.080
##
##
##                   Coef   S.E.   t    Pr(>|t|)
##  Intercept        9.3581 6.5354 1.43 0.1573
##  breakthrough_age 1.0625 0.2220 4.79 <0.0001
##
```

Simple linear regression with one categorical predictor

The next model we are going to fit is for `first_award_age` on `traditional_gender`. Once again, Specify the model, using proper LaTex, below:

$$E(\text{first award age}|\text{gender}) = \beta_0 + \beta_1(\text{gender}).$$

Our response vector (y) is the same as above, but we code categorical design matrices differently. Treating males as referrent, create the design matrix for this model now.

```
# create design matrix using data.matrix()
X2 <- stars %>% dplyr::mutate(Int = 1, Female = ifelse(gender_traditional ==
    "F", 1, 0)) %>% dplyr::select(Int, Female) %>% data.matrix()
```

Use `my_ls()` to calculate the same three values for this model, and create a table of the same format.

```
my_ls(y = y, X = X2, cap = "Modeling age at first award on gender.")
```

Check yourself. Since we want males to be our reference group, we need to recode `gender_traditional` as a factor to make sure that `ols()` will know.

	Variable	Effect	SE	Z	P
1	Intercept	42.53	2.00	21.32	<0.001
2	Gender	-11.09	3.99	-2.78	0.003

Table 2: Modeling age at first award on gender.

```r
# make gender_traditional a factor
stars %<>% mutate(gender_traditional = factor(gender_traditional, levels = c("M",
    "F")))

# fit the same model using ols()
ols(formula = first_award_age ~ gender_traditional, stars)
```

```
## Linear Regression Model
##
##  ols(formula = first_award_age ~ gender_traditional, data = stars)
##
##                 Model Likelihood     Discrimination
##                    Ratio Test           Indexes
##  Obs       63    LR chi2      7.53    R2       0.113
##  sigma13.7696    d.f.            1    R2 adj   0.098
##  d.f.      61    Pr(> chi2) 0.0061    g        4.272
##
##  Residuals
##
##      Min      1Q  Median      3Q     Max
##  -22.532  -8.032  -2.532   5.968  37.468
##
##
##                        Coef     S.E.   t     Pr(>|t|)
##  Intercept              42.5319 2.0085 21.18 <0.0001
##  gender_traditional=F -11.0944 3.9855 -2.78 0.0071
##
```

Multiple linear regression with interacting continuous and categorical predictors

Our last model seeks to incorporate separate effects of `breakthrough_age` for male and female entertainers, which we accomplish via an interaction between predictors `breakthrough_age` and `traditional_gender`. Thus, we specify the following model:

$$E(\text{first award age}|\text{breakthrough age, gender}) = \beta_0+\beta_1(\text{breakthrough age})+\beta_2(\text{gender})+\beta_3(\text{breakthrough age})\times(\text{gender}).$$

Once more, we can recycle `y` from the first model, but our design matrix `X` now needs:

1. an intercept,
2. the continuous `breakthrough_age` values,
3. the indicator for female (as in model 2), and
4. the interaction between `breakthrough_age` and the female indicator.

```r
# create design matrix using data.matrix() for model: first_award_age ~
# breakthrough_age*gender_traditional
X3 <- stars %>% dplyr::mutate(Int = 1, Female = ifelse(gender_traditional ==
    "F", 1, 0), FemaleBreakthroughAge = Female * breakthrough_age) %>% dplyr::select(Int,
    breakthrough_age, Female, FemaleBreakthroughAge) %>% data.matrix()
```

Use `my_ls()` again to calculate the same three values for this model, and create a table of the same format.

```
my_ls(y = y, X = X3, cap = "Modeling age at first award on breakthrough age, controlling for gender and
```

	Variable	Effect	SE	Z	P
1	Intercept	15.43	7.12	2.17	0.015
2	Gender	0.91	0.00	Inf	<0.001
3	Intercept	-26.32	28.16	-0.94	0.175
4	Gender	0.76	1.10	0.69	0.244

Table 3: Modeling age at first award on breakthrough age, controlling for gender and the interaction.

And check yourself against `rms::ols()`.

```
# fit the same model using ols()
ols(formula = first_award_age ~ breakthrough_age * gender_traditional, data = stars)
```

```
## Linear Regression Model
##
##  ols(formula = first_award_age ~ breakthrough_age * gender_traditional,
##      data = stars)
##
##                  Model Likelihood     Discrimination
##                     Ratio Test           Indexes
##  Obs       63    LR chi2     24.26    R2       0.320
##  sigma12.2608    d.f.            3    R2 adj   0.285
##  d.f.      59    Pr(> chi2) 0.0000    g        9.292
##
##  Residuals
##
##      Min      1Q  Median      3Q     Max
##  -15.517  -8.693  -3.587   5.044  35.743
##
##
##                                         Coef     S.E.    t     Pr(>|t|)
##  Intercept                               15.4260  7.1153  2.17 0.0342
##  breakthrough_age                         0.9132  0.2320  3.94 0.0002
##  gender_traditional=F                   -26.3204 28.1571 -0.93 0.3537
##  breakthrough_age * gender_traditional=F   0.7591  1.0943  0.69 0.4906
##
```

Predicting based on models

Manually calculating predictions

Begin by plotting `breakthrough_age` against `first_award_age`.

```
# make a scatterplot of breakthrough_age vs. first_award_age
mod1_plot <- stars %>% ggplot(aes(x = breakthrough_age, y = first_award_age)) +
    geom_point()
```

Manually calculate the predicted values for `first_award_age` based on your `beta_hat1` estimates for `breakthrough_age` from 10 to 55 years old.

```r
# create design matrix to predict age at first award for breakthrough ages
# 10-55
X_pred <- data.frame(Int = 1, breakthrough_age = seq(from = 10, to = 55, by = 1)) %>%
    data.matrix()

# use matrix operations to predict y
y_pred <- X_pred %*% beta_hat1
```

Add a line to your `mod1_plot` with the predicted values from the model.

```r
# create a dataframe with columns for x and predicted y
pred_df <- data.frame(x = seq(from = 10, to = 55, by = 1), y = y_pred)

# add geom_line() layer to mod1_plot with the model predictions for
# first_award_age
mod1_plot <- mod1_plot + geom_line(data = pred_df, aes(x = x, y = y), col = "blue",
    lty = 2)
```

Lastly, let's add the model equation to the plot (because we're fancy). Before we put it into the plot, let's make sure we're comfortable with `paste()`.

```r
# test your paste() statement to make sure you like how it looks
my_mod_eq <- paste(beta_hat1["Int", 1], " + ", beta_hat1["breakthrough_age",
    1], "x", sep = "")

# now add an annotate() layer to mod1_plot with the model equation
mod1_plot <- mod1_plot + annotate(geom = "text", x = 45, y = 72, label = my_mod_eq)

# and print your plot
mod1_plot
```

Invariance of predicted values to linear transformation

Often, we like to transform predictors so that the coefficients can be interpreted more meaningfully. Say we're more interested in 5 year changes in `breakthrough_age` than 1 year changes. Let's transform `breakthrough_age`, refit model 1 (recall: `first_award_age ~ breakthrough_age`), and compare predicted values between the original and linearly transformed models.

```r
# add a column for breakthrough_age_centered by subtracting the median
stars %<>% dplyr::mutate(breakthrough_age_5yrs = breakthrough_age/5)

# create the design matrix again
X4 <- stars %>% dplyr::mutate(Int = 1) %>% dplyr::select(Int, breakthrough_age_5yrs) %>%
    data.matrix()

# use matrix operations to calculate least-squares estimates beta0, beta1
beta_hat1_5yrs <- solve(t(X4) %*% X4) %*% t(X4) %*% y %>% round(digits = 3)

# create design matrix to predict age at first award for breakthrough ages
# 10-55
X_pred <- data.frame(Int = 1, breakthrough_age_5yrs = seq(from = 10, to = 55,
    by = 1)/5) %>% data.matrix()

# and predict first_award_age for ages 10-55 again using matrix operations
```

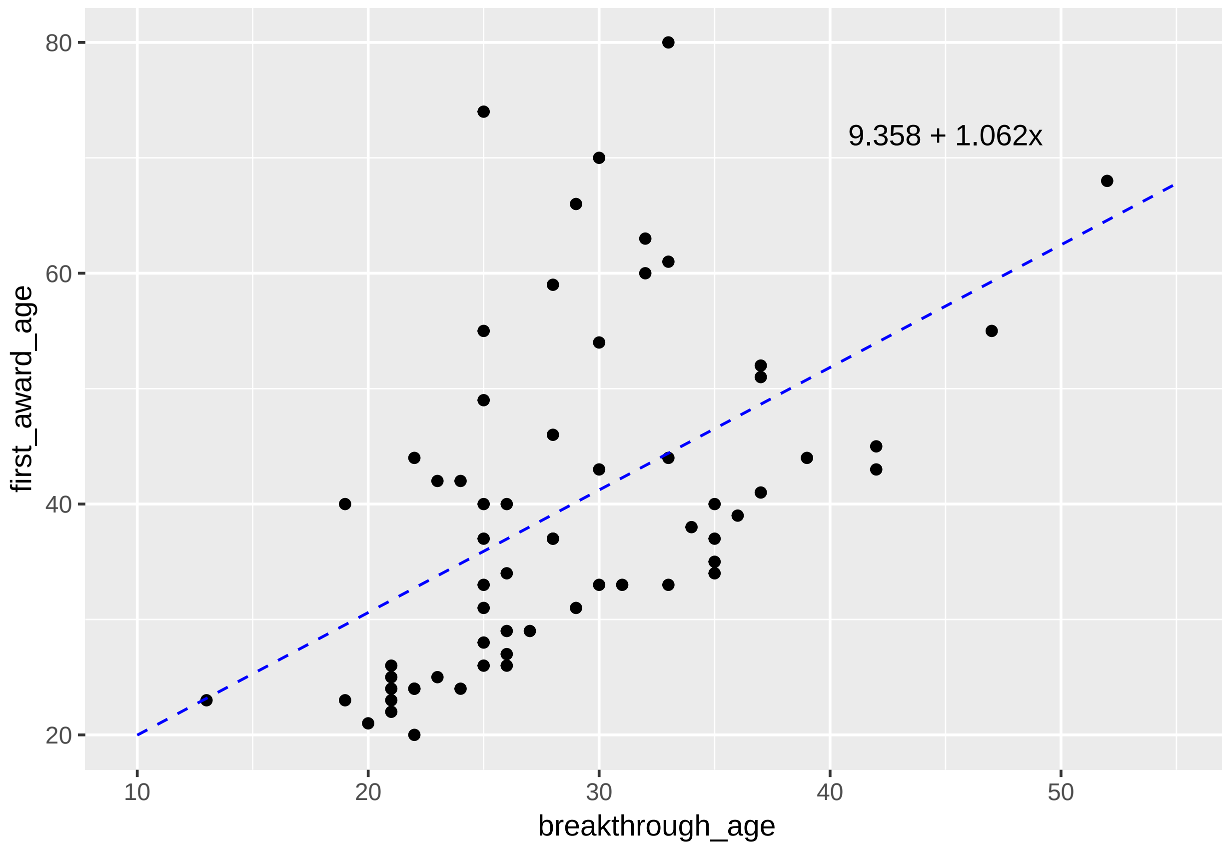

Figure 1: Modeling age at first award on breakthrough age using OLS regression.

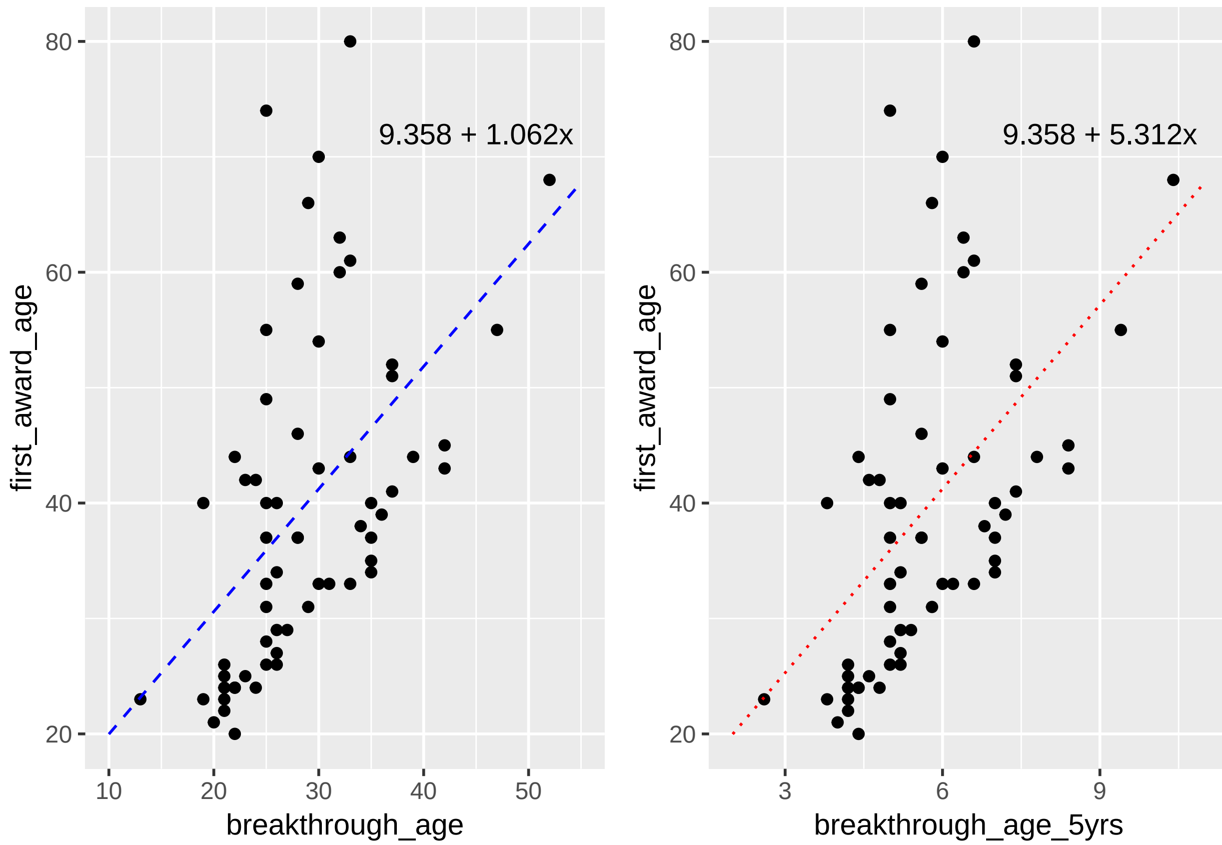

Figure 2: Illustrating prediction invariance to linear transformations in the predictor.

```
y_pred_5yrs <- X_pred %*% beta_hat1_5yrs

# create a dataframe with columns for x and predicted y
pred_df <- data.frame(x = seq(from = 10, to = 55, by = 1)/5, y = y_pred_5yrs)
```

Create the same plot as before with `breakthrough_age_5yrs` against `first_award_age`, overlaid with your transformed model predictions. Use `ggpubr:ggarrange()` to print this plot side-by-side with the original.

```
# create the same plot using the transformed model test your paste()
# statement to make sure you like how it looks
my_mod_eq <- paste(beta_hat1_5yrs["Int", 1], " + ", beta_hat1_5yrs["breakthrough_age_5yrs",
    1], "x", sep = "")

mod1_plot_5yrs <- stars %>% ggplot(aes(x = breakthrough_age_5yrs, y = first_award_age)) +
    geom_point() + geom_line(data = pred_df, aes(x = x, y = y), col = "red",
    lty = 3) + annotate(geom = "text", x = 9, y = 72, label = my_mod_eq)
# print this plot next to the original
ggarrange(mod1_plot, mod1_plot_5yrs)
```

This is follows from **Rencher Theorem 7.3e.**

Full rank transformation on predictor

Take any full-rank 2×2 matrix $\boldsymbol{K}_1$ (e.g. sampling the elements independently from a Normal distribution).

```r
# generate full-rank transformation matrix K
K <- matrix(data = rnorm(2 * 2, 5, 2), ncol = 2)
```

Just for fun, let's make sure that this matrix is full-rank.

You: "But Sarah, didn't we just design it to be?"

Sarah: "Yes, but wouldn't you like to know how to calculate the rank of a matrix in R anyway?"

```r
# check Rank(K)
qr(K)$rank
```

```
## [1] 2
```

There are some other helpful matrix operations in `R` that I would recommend for checking homework, spending a Friday night at home, etc. that can be found here.

Now, consider a model adjusting for `breakthrough_age` and `tranditional_gender` without the interaction. Construct this design matrix,

```r
# create the design matrix for model: first_award_age ~ breakthrough_age +
# traditional_gender
X5 <- stars %>% dplyr::mutate(Int = 1, Female = ifelse(gender_traditional ==
    "F", 1, 0)) %>% dplyr::select(Int, breakthrough_age, Female) %>% data.matrix()
```

and fit this model.

```r
# fit the model for y ~ X5
beta_hat5 <- solve(t(X5) %*% X5) %*% t(X5) %*% y
```

Next, let's do a full-rank linear transformation of our design matrix `X5`.

```r
# full-rank linear transformation of X5 be careful: should you be
# transforming the intercept column?
Z <- cbind(X5[, 1], X5[, -1] %*% K)
```

What do we know about Rank($\boldsymbol{Z}$)? First, check that `X5` is full-rank.

```r
# check rank(X5)
qr(X5)$rank
```

```
## [1] 3
```

Now, since `X5` is full-rank and `Z` is a full-rank, square matrix, we have that multiplying `X5` by `Z` does not change the rank (by Rencher Theorem 2.4.). Check that the rank of `Z` equals the rank of `X5`.

```r
# check that Rank(Z) = Rank(X5)
qr(Z)$rank == qr(X5)$rank
```

```
## [1] TRUE
```

Fit the model for `first_award_age` on the transformed design matrix `Z` now.

```r
# fit the model for y ~ Z
beta_hat5_trans <- solve(t(Z) %*% Z) %*% t(Z) %*% y
```

Last, we compare the fitted values for the 63 entertainers based on the model before and after full-rank linear transformation.

```r
# check that predicted values for the models are the same
table(round(X5 %*% beta_hat5, 6) == round(Z %*% beta_hat5_trans, 6))
```

```
##
```

```
## TRUE
##   63
```

This is Rencher Theorem 7.3e Corollary 1.