Homework 1

STA2101F 2020

Due October 1 2020 11.59 pm

Homework to be submitted through Quercus

You can submit this HW in Word, Latex, or R Markdown, but in future please use R Markdown. If you are using Word or Latex with a R script for the computational work, then this R script should be provided as an Appendix. In the document itself you would just include properly formatted output.

You are welcome to discuss the questions with others, but the solutions and code must be written independently. Any R output that is included in a solution should be formatted as part of the discussion (i.e. not cut and pasted from the Console).

1. **Chooose this question or the next** Find an article about the results of a study, in a scientific journal on a topic of interest to you. The article should discuss a single study, and should provide enough information on the study methods to answer the questions below.
   (a) Give the complete bibliographic reference, as well as a web link, to the published paper.
   (b) Was the study observational or a designed experiment?
   (c) What was the study population? What is the population of interest for the research?
   (d) If the study was observational, was it a prospective, or a retrospective study? If it was an experiment, was it randomized?
   (e) What were the units of analysis?
   (f) What was the primary endpoint and the main analysis of this endpoint?
   (g) What were the main conclusions of the study, in your own words?
2. **Choose this question or the previous** A short article by Professor Rob Hyndman in March describes two approaches to forecasting, time series modelling and agent-based modelling. In the latest release from the Public Health Agency of Canada, there are two forecasts, one on slide 10 and one on slide 12. The government has been criticized for not releasing details of its models, and the latest slide deck does have some references to the literature. In particular on slide 12 they link to this preprint.
   (a) Are these forecasts on slide 10 and slide 12 based on time series modelling or

agent-based modelling?

(b) Did the paper mentioned on slide 12 compare its forecasts to data? What data source(s) did they use?

(c) How many compartments does their SEIR model contain?

(d) They refer on p.15 at the beginning of Section 4 to a *fraction of contact reduction*, and its *critical threshold of the model*. What does this mean, and what is the critical threshold?

(e) How do you think PHAC constructed the forecast on Slide 12?

3. Suppose we have n measurements on a a response y, two explanatory variables x and z, and we assume a model of the form

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \epsilon_i, \quad i = 1, \ldots, n.$$

Further, assume that x is a continuous variate, but $z_i = \pm 1$; for example, x might be age, and z might be gender. We will also make the usual assumptions that the observations are independent, and that $\epsilon_i \sim (0, \sigma^2)$.

(a) Express this model as $y = X\beta + \epsilon$, where y and ϵ are $n \times 1$ vectors, making explicit the entries of the X matrix.

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & z_1 & x_1 z_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & z_n & x_n z_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

(b) Show that $E(y_i \mid x_i, z_i = -1)$ and $E(y_i \mid x_i, z_i = +1)$ are lines, and give their slopes and intercepts.

The respective expected values are $\beta_0 + \beta_1 x_i - \beta_2 - \beta_3 x_i$ and $\beta_0 + \beta_1 x_i + \beta_2 + \beta_3 x_i$. They have intercepts $\beta_0 - \beta_2, \beta_0 + \beta_2$, respectively, and slopes $\beta_1 - \beta_3, \beta_1 + \beta_3$, respectively.

(c) What conclusions are implied by the hypothesis $\beta_3 = 0$? What conclusions are implied by the hypothesis $\beta_2 = 0$? What conclusions are implied by the hypothesis $\beta_3 = \beta_2 = 0$?

If $\beta_3 = 0$ the two lines (males and females) are parallel; if $\beta_2 = 0$ they have the same intercept, and if both are zero they are the same line.

(d) Give an explicit expression for the least squares estimate of β_2, under the simpler model that assumes $\beta_3 = 0$.

Assume without loss of generality that $\sum x_i = 0$ and $\sum z_i = 0$. (If not, replace x_i with $x_i' = x_i - \bar{x}$, and similarly for z, and proceed.) Then $\hat{\beta}_0 = \bar{y}$, and the equations for $\hat{\beta}_1$ and $\hat{\beta}_2$ are

$$\sum x_i y_i - \hat{\beta}_1 \sum x_i^2 - \hat{\beta}_2 \sum z_i x_i = 0 \quad (1)$$

$$\sum z_i y_i - \hat{\beta}_1 \sum x_i z_i - \hat{\beta}_2 \sum z_i^2 = 0. \quad (2)$$

Using (2), we have

$$\hat{\beta}_2 = \frac{\sum z_i y_i}{\sum z_i^2} - \hat{\beta}_1 \frac{\sum z_i x_i}{\sum z_i^2},$$

which is the regression of y on z, corrected for the regression of x on z, with a factor of $\hat{\beta}_1$. The expression for $\hat{\beta}_1$ is messier, but eventually can be expressed as

$$\hat{\beta}_1 = \frac{S_{xy} - S_{zy}S_{xz}/S_{zz}}{S_{xx} - S_{xz}^2/S_{zz}},$$

where the shorthand $S_{xy} = \sum x_i y_i$, $S_{xx} = \sum x_i^2$ and so on has been used (note that the results lead to $\hat{\beta}_2 = \frac{S_{zy} - S_{xy}S_{xz}/S_{xx}}{S_{zz} - S_{xz}^2/S_{xx}}$). In the older literature, with z regarded as a factor variable indexing two treatments, this is called the analysis of covariance. S_{zy} is the difference between the mean of y under the high level of the factor ($z = 1$) and the mean under the low level ($z = -1$), so represents the treatment effect. The estimate $\hat{\beta}_2$ is an adjusted treatment difference, adjusted for the covariate x.

4. The dataset `teengamb` concerns a study on teenage gambling in Britain. You can get more information about the data with `?teengamb`, but you will first need `library(faraway);data(teengamb)`, and possibly `install.packages`("faraway"). The questions below are adapted from FLM.

   (a) Using `gamble` as the response, fit a linear regression model to the other four variables and give a table of the estimated coefficients and their estimated standard errors.

   The estimated coefficients and their estimated standard errors for the full model with all four predictors are given by the following:

	Estimate	Standard Error
Intercept	22.5556506	17.1968034
sex	-22.1183301	8.2111145
status	0.0522338	0.2811115
income	4.9619792	1.0253923
verbal	-2.9594935	2.1721503

   (b) With other explanatory variables held fixed, what is the estimated difference in expenditure on gambling for males compared to females?

   When all variables except sex is held fixed, the estimated difference in the mean expenditure on gambling for males compared to females is:

$$\begin{aligned}
&(\hat{\beta}_{\text{sex}} \cdot 0 + \hat{\beta}_{\text{status}} \cdot x_{\text{status}} + \hat{\beta}_{\text{income}} \cdot x_{\text{income}} + \hat{\beta}_{\text{verbal}} \cdot x_{\text{verbal}}) \\
&- (\hat{\beta}_{\text{sex}} \cdot 1 + \hat{\beta}_{\text{status}} \cdot x_{\text{status}} + \hat{\beta}_{\text{income}} \cdot x_{\text{income}} + \hat{\beta}_{\text{verbal}} \cdot x_{\text{verbal}}) \\
&= 22.11833
\end{aligned}$$

(c) Fit a model with just `income` as a predictor, and use an F-test to compare it to the full model.

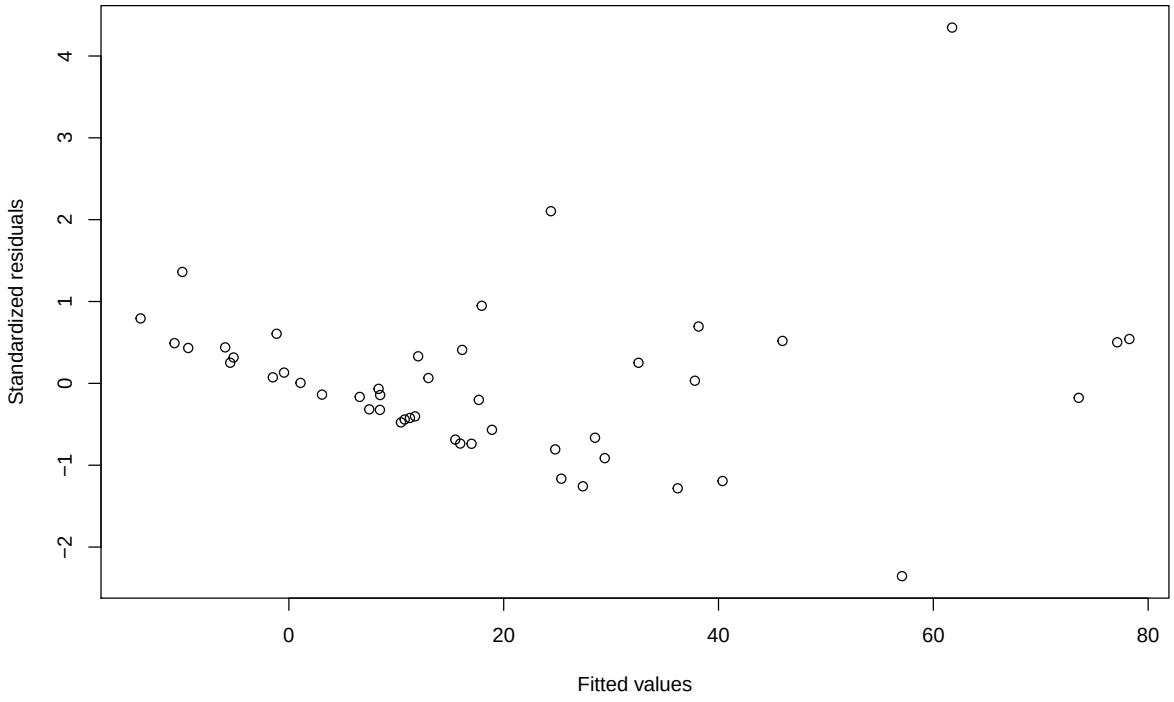
Our hypothesis for the F-test is the following:

$$H_0 : \beta_{\text{sex}} = \beta_{\text{status}} = \beta_{\text{verbal}} = 0$$
$$H_A : \beta_{\text{sex}} \neq 0, \beta_{\text{status}} \neq 0 \text{ or } \beta_{\text{verbal}} \neq 0$$

Since the p-values is 0.012, we have evidence that the H_0 is not supported by the data. In other words, at least one of the predictors 'sex', 'status' or 'verbal' has a significant effect when the predictor 'income' is in the model.

(d) Do the usual linear model assumptions appear to be satisfied for this data? Why or why not? Include at most two plots to support your answer.

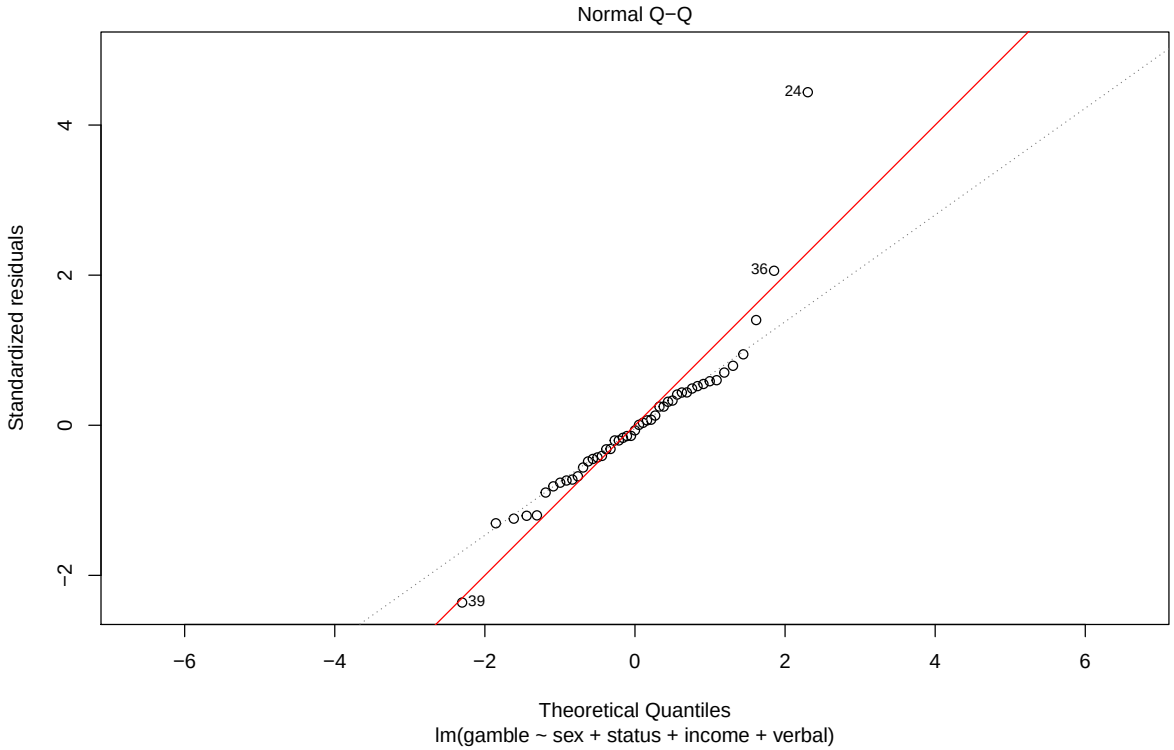
The Gauss-Markov theorem requires the error terms to be homoscedastic (i.e. equal variance) and uncorrelated between each other. The heteroscedasticity can be assessed visually by plotting the fitted values against the standardized residuals as follows:



Here, the variance of the standardized residuals does appear to increase when the fitted values increase showing a sign of heterscedasticity (however, this visual inspection must be followed by a formal test for equal variance e.g. Levene's test). Correlated error will be discussed later in the course.

In addition to the Gauss-Markov assumption, statistical tests such as the F-test above assumes the error terms to follow a Gaussian distribution. A severe departure from normality can affect the type I error and power of these tests that assume normality. A

QQ-plot can reveal a departure from normality. The QQ-plot for our model is shown by the following figure:



When the standardized residuals follow approximately a Gaussian distribution, the actual quantiles and the theoretical quantiles will form a pattern around the diagonal line in red. Here, an S-shaped pattern can be observed, indicating a violation of normality. However, one should not conclude based on the visual inspection. A formal test for normality (e.g. Shapiro–Wilk test) is required to support this claim.

In addition to the assumptions above, one must keep in mind that a linearity in the functional form of the predictors was assumed. This, however, might not be true and it does not to appear to be the case, here. In the plot of fitted values against the standardized residuals above, a pattern could be observed with a visual inspection. If the linearity in the predictors was true, residuals do not form any pattern.

(e) Predict the amount that a male with `status`, `income` and `verbal` at the maximum values in this data set would gamble, along with a 95% prediction interval.

The following are the maximum values for 'status', 'income' and 'verbal' for a male:

sex	status	income	verbal
0	75	15	10

For this data point, a predicted expenditure on gambling is 71.30794 with a 95% prediction interval of [17.06588, 125.55].

(f) Fit a model with `sqrt(gamble)` as the response but with the same explanatory

variables. Give a 95% prediction interval for the individual in (e), taking care to convert the interval to the original units of the response.

Using a square root transformtion, the predicted square root expenditure on gambling is 8.697723 with a 95% prediction interval of [3.715767, 13.67968]. Back transforming to the original scale, the predicted expenditure is 75.65038 with a 95% prediction interval of [13.80692, 187.1336].

(g) A model selection strategy that is easily applied here is all possible subsets. Assuming `sex` must be included in all the models, fit all possible combinations of covariates. How stable is the estimated effect of `sex` across these models?

The estimated effect, standard error, observed T-test statistic and the corresponding p-value of 'sex' across these models are as follows:

	Estimate	Standard Error	T-test statistic	p-value
sex	-25.90921	8.647659	-2.996095	0.0044366
sex + status	-35.70937	9.489859	-3.762898	0.0004934
sex + income	-21.63439	6.808797	-3.177417	0.0027173
sex + verbal	-27.72208	8.416655	-3.293717	0.0019570
sex + status + income	-24.33934	8.127413	-2.994722	0.0045427
sex + status + verbal	-33.75202	9.683931	-3.485363	0.0011444
sex + income + verbal	-22.96022	6.770575	-3.391177	0.0015024
sex + status + income + verbal	-22.11833	8.211115	-2.693706	0.0101118

In all the models above, test statistics are large, leading us to conclude that the effects of sex is significant.

5. (SM Exercise 8.1.1) Which of the following can be written as linear regression models, (i) as they are, (ii) when a single parameter is held fixed, (iii) after transformation? For those that can be so written, give the response variable and the form of the design matrix.

(a) $y = \beta_0 + \beta_1/x + \beta_2/x^2 + \epsilon$

This is a linear regression since it's linear in β's. The design matrix is the following:

$$\begin{bmatrix} 1 & 1/x_1 & 1/x_1^2 \\ & \vdots & \\ 1 & 1/x_n & 1/x_n^2 \end{bmatrix}$$

(b) $y = \beta_0/(1 + \beta_1 x) + \epsilon$

This is non-linear in β's. However, if β_1 is held fixed, the model is linear with the following design matrix:

$$\begin{bmatrix} \frac{1}{1+\beta_1 x_1} \\ \vdots \\ \frac{1}{1+\beta_1 x_1} \end{bmatrix}$$

(c) $y = 1/(\beta_0 + \beta_1 x + \epsilon)$

This is non-linear. However, inverting both sides, the model becomes $y^{-1} = \beta_0 + \beta_1 x + \epsilon$. The design matrix is the following:

$$\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

The response is the following vector:

$$\begin{bmatrix} y_1^{-1} \\ \vdots \\ y_n^{-1} \end{bmatrix}$$

(d) $y = \beta_0 + \beta_1 x^{\beta_2} + \epsilon$

This is non-linear. However, it becomes linear by fixing β_2. The design matrix is the following:

$$\begin{bmatrix} 1 & x_1^{\beta_2} \\ \vdots & \vdots \\ 1 & x_n^{\beta_2} \end{bmatrix}$$

(e) $y = \beta_0 + \beta_1 x_1^{\beta_2} + \beta_3 x_2^{\beta_4} + \epsilon$

This is non-linear. Fixing a parameter or transforming the model will not result in a linear model.

6. **Bonus Question: highly recommended for PhD** Suppose our model for the expected value of Y is nonlinear in β:

$$y_i = \eta_i(\beta) + \epsilon_i, \quad i = 1, \ldots, n; \quad \beta \in \mathbb{R}^p$$

where we assume $\eta(\cdot)$ is a known function of β (and possibly some covariates), and ϵ_i are i.i.d. $N(0, \sigma^2)$. Show that the least squares estimate of β solves the equation

$$\{y - \eta(\beta)\}^T X(\beta) = 0,$$

where η is $n \times 1$ and X is $n \times p$. Give an expression for the (i, j)th entry of $X(\beta)$. Suggest an iterative method of solving this equation from some starting point β_0. Try this out on Example 10.1 in SM, using the nonlinear model suggested there:

$$y = \beta_0\{1 - \exp(-x/\beta_1)\} + \epsilon.$$

The objective function to minimize for the least square method is the following:

$$f(\beta) = (y - \eta(\beta))^T (y - \eta(\beta))$$

Suppose $\eta(\cdot)$ is such that $\underset{\beta}{\operatorname{argmin}} f(\beta)$ yields a unique solution. To solve for the minimum, we set the gradient of $f(\beta)$ to 0 as follows:

$$\frac{\partial f(\beta)}{\partial \beta} = (y - \eta(\beta))^T \frac{\partial \eta(\beta)}{\partial \beta} = 0$$

where the entry (i, j) of $X(\beta) = \frac{\partial \eta(\beta)}{\partial \beta}$ is $\frac{\partial \eta_i(\beta)}{\partial \beta_j}$.

A closed-form solution for the equation above may not be available. In this case, numerical methods such as gradient descent or Newton-Raphson method can be employed (note that these work when β is unconstrained; for constrained optimization, see, for example, Convex optimization by Boyd & Vandenberghe).

For the example 10.1 in SM, an implementation of Gauss-Newton method is available as follows, thanks to **Dayi Li**'s excellent work:

```
# eta function and its derivative
eta <- function(x, beta){
  beta[1]*(1 - exp(-x/beta[2]))
}

deta <- function(x, beta){
  cbind(1-exp(-x/beta[2]), -beta[1]*x/beta[2]^2*exp(-x/beta[2]))
}

# data
x <- rep(c(0.45, 1.3, 2.4, 4, 6.1, 8.05, 11.15, 13.15, 15), each=3)
y <- c(0.34170, -0.00438, 0.82531, 1.77967, 0.95384,
       0.64080, 1.75136, 1.27497, 1.17332, 3.12273,
       2.60958, 2.57429, 3.17881, 3.00782, 2.67061,
       3.05959, 3.94321, 3.43726, 4.80735, 3.35583,
       2.78309, 5.13825, 4.70274, 4.25702, 3.60407,
       4.15029, 3.42484)

# Newton-Raphson method
N.M <- function(beta0, x, y, epsilon){
```

```r
  beta <- beta0
  r2 <- 1
  while (r2 > epsilon) {
    dy <- y - eta(x, beta)
    J <- deta(x, beta)
    db <- solve(t(J)%*%J)%*%(t(J)%*%dy)
    beta <- beta + db
    r2 <- abs((y - eta(x,beta))%*%(y-eta(x,beta))/r2-1)
  }
  return(beta)
}

# Result
beta <- N.M(c(0.4,0.5), x, y, 0.000001)
```

Q4.

```
> library(faraway)
> data(teengamb)
>
> fit_full <- lm(gamble ~ sex + status + income + verbal, data=teengamb)
> fit_full_tb <- summary(fit_full)$coefficients[,1:2]
>
> fit1 <- lm(gamble ~ income, data=teengamb)
> anova(fit1, fit_full)
Analysis of Variance Table

Model 1: gamble ~ income
Model 2: gamble ~ sex + status + income + verbal
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     45 28009
2     42 21624  3    6384.8 4.1338 0.01177 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> plot(fit_full$fitted.values, scale(fit_full$residuals), xlab="Fitted values",
ylab="Standardized residuals")
>
> plot(fit_full, which = 2, asp=1, main="")
> abline(coef=c(0, 1), col='red')
>
> x <- data.frame(sex=0, status=max(teengamb$status[teengamb$sex==0]),
income=max(teengamb$income[teengamb$sex==0]), verbal=max(teengamb$verbal[teengamb$sex==0
```

```
> predict(fit_full, x, interval="prediction")
       fit      lwr    upr
1 71.30794 17.06588 125.55
>
> teengamb$gamble_sqrt <- sqrt(teengamb$gamble)
> fit_full_sqrt <- lm(gamble_sqrt ~ sex + status + income + verbal, data=teengamb)
> predict(fit_full_sqrt, x, interval="prediction")
       fit      lwr      upr
1 8.697723 3.715767 13.67968
> predict(fit_full_sqrt, x, interval="prediction")^2
       fit      lwr      upr
1 75.65038 13.80692 187.1336
>
>       tb <- rbind(summary(lm(gamble ~ sex, data=teengamb))$coefficients['sex', 1:2],
+       summary(lm(gamble ~ sex + status, data=teengamb))$coefficients['sex', 1:2],
+       summary(lm(gamble ~ sex + income, data=teengamb))$coefficients['sex', 1:2],
+       summary(lm(gamble ~ sex + verbal, data=teengamb))$coefficients['sex', 1:2],
+       summary(lm(gamble ~ sex + status + income, data=teengamb))$coefficients['sex', 1
+       summary(lm(gamble ~ sex + status + verbal, data=teengamb))$coefficients['sex', 1
+       summary(lm(gamble ~ sex + income + verbal, data=teengamb))$coefficients['sex', 1
+       summary(lm(gamble ~ sex + status + income + verbal, data=teengamb))$coefficients
>
>       colnames(tb)[2] <- "Standard Error"
>       rownames(tb) <- c("sex", "sex + status", "sex + income", "sex + verbal", "sex +
> tb
                               Estimate Standard Error
sex                            -25.90921       8.647659
sex + status                   -35.70937       9.489859
sex + income                   -21.63439       6.808797
sex + verbal                   -27.72208       8.416655
sex + status + income          -24.33934       8.127412
sex + status + verbal          -33.75202       9.683931
sex + income + verbal          -22.96022       6.770575
sex + status + income + verbal -22.11833       8.211115
```