TEXAS A&M UNIVERSITY

Department of Statistics

STAT 608 - Regression Analysis

Homework V

Salih Kilicli

July 6, 2019

Question 1: **Is the following statement true or false? If you believe that the statement is false, provide a brief explanation.**

"Suppose that a straight line regression model has been fit to bivariate data set of the form $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Furthermore, suppose that the distribution of X appears to be normal while the Y variable is highly skewed. A plot of standardized residuals from the least squares regression line produce a quadratic pattern with increasing variance when plotted against $(x_1, x_2, \ldots, x_n)$. In this case, one should consider adding a quadratic term in X to the regression model and thus consider a model of the form $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$."

Solution: No, even though it is a reasonable model, quadratic model doesn't fix the highly skewed Y variable issue. In order to deal with this issue, one can check an inverse response plot and transform only the response variable using Box-Cox method, since the transformed version of Y will be normaly distributed. Moreover, depending on the extend of the increase in the variance, a log transformation might be needed to reduce the increase in the error variance.

Question 2: **(Section 3.3.1) Let X be a binomial (n, p) random variable and set $Y = X/n$. Find a function f for which $Var[f(Y)]$ does not depend upon p.**

Solution : Since $X \sim B(n, p),$ and $Y = X/n$:

$$E[X] = np \Longrightarrow E[Y] = \frac{1}{n}E[X] = p$$
$$Var(X) = np(1-p) \Longrightarrow Var(Y) = \frac{1}{n^2}E(X) = \frac{p(1-p)}{n}$$

Expanding $f(Y)$ into Taylor series (linear approximation) in terms of $f(E[Y])$ and taking variance of the both sides yields;

$$Var[f(Y)] \approx [f'(E[Y])]^2 Var(Y)$$

Plugging the values of $E[Y]$ and $Var(Y)$ yields:

$$Var[f(Y)] \approx [f'(p)]^2 \, \frac{p(1-p)}{n}$$

Since $Var[f(Y)]$ is requested to be independent of p this implies $[f'(p)]^2$ should at least include a factor of $p(1-p)$ in the denominator. Let's consider the simplest case and assume $[f'(p)] = \dfrac{1}{p(1-p)}$ Then, taking square root of and then integrating both sides of the equation yields: (Consider the positive sign for simplicity)

$$\int f'(p)dp = \int \frac{dp}{\sqrt{p(1-p)}} \quad [\text{Substitute } p = sin^2(x), 1-p = cos^2(x)]$$
$$= \int \frac{2sin(x)cos(x)}{\sqrt{sin^2(x)cos^2(x)}}dx \quad [dp = 2sin(x)cos(x)dx]$$
$$= \int 2dx = 2x + C = 2arcsin(\sqrt{p}) + C$$

where C is a constant of integration. Notice, since $p = sin^2(x) \Longrightarrow x = arcsin(\pm\sqrt{p})$ (Again, we considered positive sign for simplicity.) Therefore, $f(Y) = 2arcsin(\sqrt{Y})$ ($C = 0$ case) is one of the candidates for which $Var(f(Y)) = 1/n$ and it does <u>not</u> depend on p.

Question 3: **In the simple linear regression model**

$$y_j = \beta_0 + \beta_1 x_j + e_j, \quad j = 1, 2, \ldots, n$$

the predicted values are defined by $\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_j$ where $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the least squares estimators of β_0 and β_1.

3.1 Show that the mean of the $y-$ values, $\bar{y}$, equals the mean of the predicted values, $\bar{\hat{y}}$.

3.2 Show that $SSreg = \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})$ where $\bar{\hat{y}} = n^{-1}\sum_{i=1}^{n}\hat{y}_i$.

3.2 Hence, show that the statistic $R^2 = \dfrac{SSreg}{SST}$ equals the square of the Pearson correlation coefficient between the pairs $(y_j, \hat{y}_j),\ j = 1, 2, \ldots, n.$

Solution: 3.1 $\bar{\hat{y}} = \dfrac{1}{n}\sum_{j=1}^{n}\hat{y}_j = \dfrac{1}{n}\sum_{j=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_j) = \hat{\beta}_0 + \hat{\beta}_1\bar{x} = (\bar{y} - \hat{\beta}_1\bar{x}) + \hat{\beta}_1\bar{x} = \bar{y}$ by (2.3).

3.2

$$\begin{aligned} SSreg &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})\{(\hat{y}_i - y_i) + (y_i - \bar{y})\} \\ &= \cancelto{0}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(\hat{y}_i - y_i)} + \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y}) \\ &= \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y}) \end{aligned}$$

since the first sum is zero. Below is the proof of why the first sum is zero:

$$\begin{aligned} \sum_{i=1}^{n}(\hat{y}_i - \bar{y})(\hat{y}_i - y_i) &= \sum_{i=1}^{n}(\cancel{\hat{\beta}_0} + \hat{\beta}_1 x_i - (\cancel{\hat{\beta}_0} + \hat{\beta}_1\bar{x}))(\hat{y}_i - y_i) \\ &= \hat{\beta}_1\sum_{i=1}^{n}(x_i - \bar{x})(-\hat{e}_i) \\ &= \hat{\beta}_1\bar{x}\cancelto{0}{\sum_{i=1}^{n}\hat{e}_i} - \hat{\beta}_1\cancelto{0}{\sum_{i=1}^{n}x_i\hat{e}_i} = 0 \end{aligned}$$

Here, notice that;

$$\begin{aligned} \sum_{i=1}^{n}\hat{e}_i &= \sum_{i=1}^{n}(y_i - \hat{y}) = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ &= \sum_{i=1}^{n}(y_i - \bar{y} + \hat{\beta}_1\bar{x} - \hat{\beta}_1 x_i) = \sum_{i=1}^{n}(y_i - \bar{y}) - \hat{\beta}_1\sum_{i=1}^{n}(x_i - \bar{x}) \\ &= \left[\sum_{i=1}^{n}y_i - n\bar{y}\right] - \hat{\beta}_1\left[\sum_{i=1}^{n}x_i - n\bar{x}\right] = \Big(n\bar{y} - n\bar{y}\Big) - \hat{\beta}_1\Big(n\bar{x} - n\bar{x}\Big) = 0 \end{aligned}$$

Moreover,

$$\sum_{i=1}^{n} x_i \hat{e}_i = \sum_{i=1}^{n} (x_i y_i - x_i \bar{y} + \hat{\beta}_1 x_i \bar{x} - \hat{\beta}_1 x_i^2) = \sum_{i=1}^{n} (x_i y_i - \bar{x} y_i) - \hat{\beta}_1 \sum_{i=1}^{n} (x_i^2 - x_i \bar{x}) \quad SXX$$

$$= \sum_{i=1}^{n} (x_i - \bar{x}) y_i - \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{SXX} SXX = \sum_{i=1}^{n} (x_i - \bar{x}) y_i - \sum_{i=1}^{n} (x_i - \bar{x}) y_i = 0$$

3.3

$$R^2 = \frac{SSreg}{SST} = \frac{\sum_{j=1}^{n} (\hat{y_j} - \bar{\hat{y}})(y_j - \bar{y})}{\sum_{j=1}^{n} (y_j - \bar{y})^2} \frac{\sum_{j=1}^{n} (\hat{y_j} - \bar{\hat{y}})^2}{\sum_{j=1}^{n} (\hat{y_j} - \bar{\hat{y}})^2} = \left[\frac{cov(y_j, \hat{y_j})}{\sigma(y_j)\sigma(\hat{y_j})} \right]^2 = \rho^2(y_j, \hat{y_j})$$

Since $\sum_{j=1}^{n} (\hat{y_j} - \bar{\hat{y}})^2 = \sum_{j=1}^{n} (\hat{y_j} - \bar{y})^2 = \sum_{j=1}^{n} (\hat{y_j} - \bar{\hat{y}})(y_j - \bar{y})$ from Question (3.2).

Question 4: **A botanist wished to develop a model to describe the relationship between the total production of photosynthetic biomass of a mesquite tree and certain easily measured aspects of the tree. The measurements that were considered relevant were**

$Y =$ **total weight (in grams) of the leaves on the tree**

$x1 =$ **canopy diameter (in meters) measured along the longest axis**

$x2 =$ **canopy diameter (in meters) measured along the shortest axis**

$x3 =$ **total height (in meters) of the mesquite tree**

$x4 =$ **canopy height (in meters) of the mesquite tree**.

The data are in the file mesquite.csv.

Solution: 4.1 Below are the summary and inverse plot results of the fitted model.

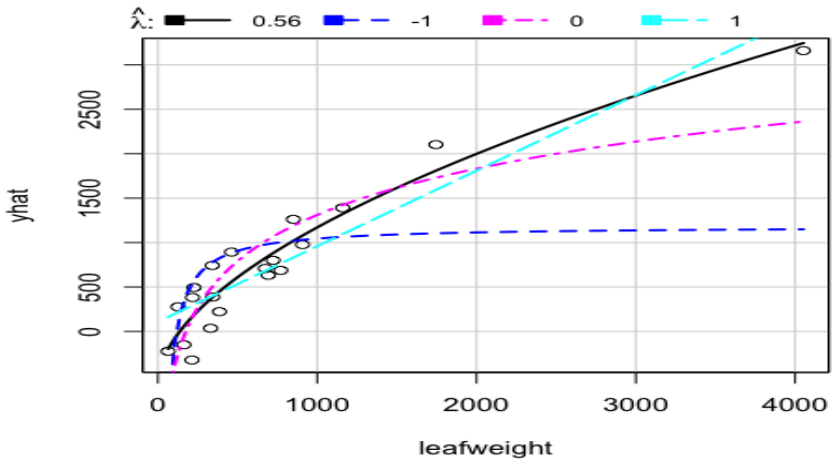
```
Call:
lm(formula = leafweight ~ Canopy.largest.diam + Canopy.shortest.diam +
    Canopy.height + Tree.hight)

Residuals:
    Min      1Q  Median      3Q     Max
-430.45 -240.13  -55.94  195.20  892.23

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1397.82     302.92  -4.614 0.000337 ***
Canopy.largest.diam    847.54     281.00   3.016 0.008681 **
Canopy.shortest.diam   -15.65     255.56  -0.061 0.951992
Canopy.height          343.94     569.42   0.604 0.554852
Tree.hight            -215.64     478.06  -0.451 0.658394
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 384.8 on 15 degrees of freedom
Multiple R-squared:  0.8506,    Adjusted R-squared:  0.8108
F-statistic: 21.35 on 4 and 15 DF,  p-value: 4.739e-06

> inverseResponsePlot(fit,key=TRUE)
      lambda     RSS
1  0.5602121 1208569
2 -1.0000000 8183808
3  0.0000000 2719565
4  1.0000000 1888750
```



Clearly, $\lambda = 1$ is not the best fit in the inverse plot, and a transformation might

be needed (Adjusted-R value doesn't look too bad). Moreover, $\lambda = 0.56$ is the closest fit and gives the smallest RSS value; therefore, $f(Y) \approx Y^{0.5}$ square root transformation might be a good option.

4.2 Below are the summary and inverse plot results of the log-model.

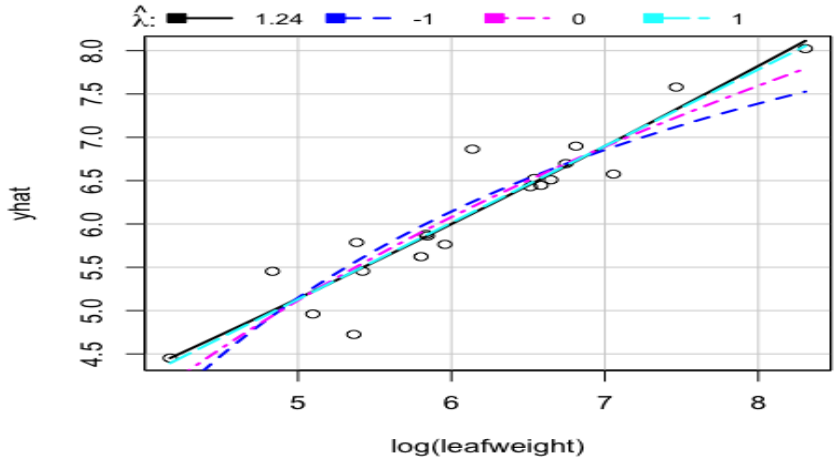
```
Call:
lm(formula = log(leafweight) ~ log(Canopy.largest.diam) + log(Canopy.shortest.diam) +
    log(Canopy.height) + log(Tree.hight))

Residuals:
     Min       1Q   Median       3Q      Max
-0.72677 -0.09396  0.03281  0.14891  0.63882

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.3043     0.3216  13.383  9.6e-10 ***
log(Canopy.largest.diam)   0.9579     0.6429   1.490   0.1569
log(Canopy.shortest.diam)  1.0194     0.4405   2.314   0.0353 *
log(Canopy.height)        -0.6040     0.7012  -0.861   0.4026
log(Tree.hight)            1.1650     0.7259   1.605   0.1294
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3685 on 15 degrees of freedom
Multiple R-squared:  0.8841,    Adjusted R-squared:  0.8532
F-statistic:  28.6 on 4 and 15 DF,  p-value: 7.307e-07

> inverseResponsePlot(fit2,key=TRUE)
     lambda      RSS
1  1.239803 1.791041
2 -1.000000 2.633677
3  0.000000 2.052683
4  1.000000 1.800726
```



Looking at the inverse plot, we see that $\lambda = 1$ is one of the best fits even though the best fit is attained for $\lambda = 1.24$. Additionally, Adjusted - R^2 value is slighlty improved and significance of the model increased (smaller p-value on F-statistics). However, looking at sqrt(standardized residuals) vs fitted values plot we can see a non-constant variance problem. Also, the number of negative coefficients compared the first model is reduced. Overall, the transformation is decent but it can be improved, for Ex. $f(Y) \approx Y^{5/4}$ transformation might be good.

4.3 We fail to reject H_0 hypothesis since $p = 0.3958 > 0.1 = \alpha$. The proof of the statement using LinearHypothesis command output in R is given below.

```
> M
     [,1] [,2] [,3] [,4] [,5]
[1,]    0    1   -1    0    0
[2,]    0    0    0    1    0
[3,]    0    0    0    0    1
> linearHypothesis(model=fit2,hypothesis.matrix=M,test="F")
Linear hypothesis test

Hypothesis:
log(Canopy.largest.diam) - log(Canopy.shortest.diam) = 0
log(Canopy.height) = 0
log(Tree.hight) = 0

Model 1: restricted model
Model 2: log(leafweight) ~ log(Canopy.largest.diam) + log(Canopy.shortest.diam) +
    log(Canopy.height) + log(Tree.hight)

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     18 2.4681
2     15 2.0368  3   0.43129 1.0587 0.3958
```

Question 5: **A group of $m+n$ arthritis sufferers was randomly divided into two groups, A and B. The patients in group A were treated with a placebo while those in group B were treated with a prescription drug. After one month of treatment, the times in seconds taken to walk 10 yards was measured. The measured times in the two groups were $A_1, \ldots, A_m$ and $B_1, \ldots, B_n$ .It is required to test the hypothesis $H0 : \mu = \eta$ where μ and η denote respectively the mean walking times in the (hypothetical) populations of sufferers treated with the placebo and prescription drug respectively. Express the data in linear model form:**

$$E[Y_i|x_i] = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, (m+n)$$

in such a manner that the hypothesis $H_0 : \mu = \eta$ is equivalent to the hypothesis $H_0 : \beta_1 = 0$. Express clearly $\hat{\beta}_0$ and $\hat{\beta}_1$ in terms of μ and η and express $Y_1, \ldots, Y_{m+n}$ in terms of the A_i and B_i. Also give the numerical values of $x_1, \ldots, x_{m+n}$.

Solution: First of all, let

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}; \quad \begin{bmatrix} x_{m+1} \\ x_{m+2} \\ \vdots \\ x_{m+n} \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{bmatrix}; \quad X = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & -1 \end{bmatrix} \text{ and } \quad Y = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_m \\ B_1 \\ \vdots \\ B_n \end{bmatrix}$$

Now, let us calculate the estimates using $\hat{\beta} = (X'X)^{-1}X'Y$.

$$X'X = \begin{bmatrix} 1 & 1 & \ldots & 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 & -1 & \ldots & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} m+n & m-n \\ m-n & m+n \end{bmatrix}$$

Therefore $det(X'X) = (m+n)^2 - (m-n)^2 = 4mn$, and the inverse matrix is:

$$(X'X)^{-1} = \frac{1}{4mn} \begin{bmatrix} m+n & -(m-n) \\ -(m-n) & m+n \end{bmatrix} = \frac{1}{4mn} \begin{bmatrix} m+n & n-m \\ n-m & m+n \end{bmatrix}$$

Now, let us multiply $(X'X)^{-1}$ by X' and then by Y to get estimates.

$$\begin{aligned} (X'X)^{-1}X' &= \frac{1}{4mn} \begin{bmatrix} m+n & n-m \\ n-m & m+n \end{bmatrix} \begin{bmatrix} 1 & 1 & \ldots & 1 & \ldots & 1 \\ 1 & 1 & \ldots & -1 & \ldots & -1 \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} \frac{1}{m} & \cdots & \frac{1}{m} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{m} & \ldots & \frac{1}{m} & \frac{-1}{n} & \ldots & \frac{-1}{n} \end{bmatrix} \end{aligned}$$

Now, $\hat{\beta} = [\hat{\beta}_0 \quad \hat{\beta}_1]'$ is given by:

$$\hat{\beta} = (X'X)^{-1}X'Y = \frac{1}{2}\begin{bmatrix} \frac{1}{m} & \cdots & \frac{1}{m} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{m} & \cdots & \frac{1}{m} & \frac{-1}{n} & \cdots & \frac{-1}{n} \end{bmatrix} \begin{bmatrix} A_1 \\ \vdots \\ A_m \\ B_1 \\ \vdots \\ B_n \end{bmatrix} = \frac{1}{2}\begin{bmatrix} \frac{1}{m}\sum_{i=1}^{m} A_i + \frac{1}{n}\sum_{i=1}^{n} B_i \\ \frac{1}{m}\sum_{i=1}^{m} A_i - \frac{1}{n}\sum_{i=1}^{n} B_i \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{2}\Big(\mu + \eta\Big) \\ \frac{1}{2}\Big(\mu - \eta\Big) \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

Therefore, $\hat{\beta}_0 = \frac{1}{2}\Big(\mu + \eta\Big)$ and $\hat{\beta}_1 = \frac{1}{2}\Big(\mu - \eta\Big)$ and clearly both of the H_0 hypotheses given above are equivalent since $\mu = \eta \Longrightarrow \hat{\beta}_1 = \frac{1}{2}\Big(\mu - \eta\Big) = 0$

Question 6: **The data in the file Hubble.csv give measurements on the distances from earth and the corresponding recession velocities of $n = 24$ nebulae. Fit a regression model**

$$E[velocity|distance] = \beta_0 + \beta_1 distance + \beta_2 (distance)^2$$

to the data and;

(a) write down the estimates of the regression coefficients;

(b) write down the correlation coefficient between the two predictor variables.

(c) The F-statistic for testing the null hypothesis that both $\hat{\beta}_1$ and $\hat{\beta}_2$ are zero is highly significant (you do not have to verify this.) However, neither of the two predictor coefficient estimates is significant at the 5% level. Describe briefly how this can be explained.

(d) Show that the issue disappears if distance is replaced by

$$distance - mean(distance) = distance - 0.911$$

in the model. Explain briefly what is going on here and how the replacement affects the interpretation of the coefficient estimate.

Solution: (a) $velocity = -63.22 + 520.80\ distance - 31.15\ distance^2$

(b) $\rho(distance, distance^2) = 0.964$ by R's cor command.

(c) Multicollinearity affects only the the independent variables that are highly correlated. Since the correlation coefficient between the two predictor variables are high, the estimates for the coefficients are unexpected (estimates, t-values and p-values are unexpectedly poor) and this results in large standard errors, wide confidence intervals, and huge confidence or prediction bands. This issue can be solved simply by subtracting the mean from the variable since it will reduce the correlation between two predictor variables.

(d) The new correlation coefficient given by R is;

$$\rho(distance - mean(distance), distance - mean(distance))^2) = 0.4744$$

Clearly, the new predictor variables are <u>not</u> highly correlated anymore (it is reduced more than 50%). Therefore, the coefficient estimates are better and more significant. Fitting the mean centered model gives the exactly same plots. Accordingly, the sum-of-squares is the same, as are results of model comparisons. Here is the comparison of the summaries of the first model and improved model:

```
Call:
lm(formula = velocity ~ distance + I(distance^2))

Residuals:
    Min      1Q  Median      3Q     Max
-410.28 -150.40  -12.89  143.94  493.56

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     -63.22     127.05  -0.498   0.6240
distance        520.80     290.15   1.795   0.0871 .
I(distance^2)   -31.15     130.75  -0.238   0.8140
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 238.1 on 21 degrees of freedom
Multiple R-squared:  0.6245,    Adjusted R-squared:  0.5888
F-statistic: 17.47 on 2 and 21 DF,  p-value: 3.411e-05
```

Figure 1: Fit with highly correlated predictors

```
Call:
lm(formula = velocity ~ (I(distance - mean) + I((distance - mean)^2)))

Residuals:
    Min      1Q  Median      3Q     Max
-410.28 -150.40  -12.89  143.94  493.56

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            385.56      71.33   5.406 2.31e-05 ***
I(distance - mean)     464.03      87.36   5.312 2.88e-05 ***
I((distance - mean)^2)  -31.15     130.75  -0.238    0.814
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 238.1 on 21 degrees of freedom
Multiple R-squared:  0.6245,    Adjusted R-squared:  0.5888
F-statistic: 17.47 on 2 and 21 DF,  p-value: 3.411e-05
```

Figure 2: Fit with mean centered predictors