550.413 Assignment 3

Fall 2016

Instruction: This assignment consists of 4 problems. The assignment is due on **Wednesday, November 2, 2016** at 3pm, in class. If you cannot make it to class, please leave the assignment under the door at Whitehead Hall 306E and email the course instructor. If possible, please type up your assignments, preferably using LaTeX.

Problem 1: (10pts)

Let $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ be a linear model where $\mathbf{X}$ is of size $n \times p$ and the error terms $\boldsymbol{\epsilon}$ are independent, normally distributed with mean 0 and variance σ^2. Suppose furthermore that the columns of $\mathbf{X}$ can be partitioned as

$$\mathbf{X} = \begin{bmatrix} \mathbf{W} & \mathbf{Z} \end{bmatrix}$$

where $\mathbf{W}$ is of size $n \times q$ and is of full-column rank and $\mathbf{Z}$ is of size $n \times (p - q)$ and is of full-column rank, for some q satisfying $1 \leq q \leq p$, and that $\mathbf{W}^T\mathbf{Z} = \mathbf{0}$. We now partition $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$ where $\boldsymbol{\beta}_1$ is of size $q \times 1$ and $\boldsymbol{\beta}_2$ is of size $(p - q) \times 1$. Let $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix}$ be the least square estimate of $\boldsymbol{\beta}$.

(a) Show that $\hat{\boldsymbol{\beta}}_1 = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\boldsymbol{y}$ and $\hat{\boldsymbol{\beta}}_2 = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{y}$.

(b) Show that $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ are independent.

(c) Let $\boldsymbol{a}$ be a $q \times 1$ vector and $\boldsymbol{b}$ be a $(q-p) \times 1$ vector. Let (l_1, u_1) and (l_2, u_2) be the individual 95% confidence intervals for $\boldsymbol{a}^T\boldsymbol{\beta}_1$ and $\boldsymbol{b}^T\boldsymbol{\beta}_2$ based on $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$, respectively. Is the confidence interval (l_1, u_1) independent of the confidence interval (l_2, u_2) ? Justify your answer.

Solution:

For part (a), we argue directly as follows

$$\begin{aligned}
\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y} &= \left(\begin{bmatrix}\mathbf{W}^T \\ \mathbf{Z}^T\end{bmatrix}\begin{bmatrix}\mathbf{W} & \mathbf{Z}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{W}^T \\ \mathbf{Z}^T\end{bmatrix}\boldsymbol{y} \\
&= \begin{bmatrix}\mathbf{W}^T\mathbf{W} & \mathbf{W}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{W} & \mathbf{Z}^T\mathbf{Z}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{W}^T\boldsymbol{y} \\ \mathbf{Z}^T\boldsymbol{y}\end{bmatrix} \\
&= \begin{bmatrix}\mathbf{W}^T\mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}^T\mathbf{Z}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{W}^T\boldsymbol{y} \\ \mathbf{Z}^T\boldsymbol{y}\end{bmatrix} \\
&= \begin{bmatrix}(\mathbf{W}^T\mathbf{W})^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{Z}^T\mathbf{Z})^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{W}^T\boldsymbol{y} \\ \mathbf{Z}^T\boldsymbol{y}\end{bmatrix} \\
&= \begin{bmatrix}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\boldsymbol{y} \\ (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{y}\end{bmatrix}
\end{aligned}$$

and hence $\hat{\boldsymbol{\beta}}_1 = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\boldsymbol{y}$ and $\hat{\boldsymbol{\beta}}_2 = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{y}$ as desired.

For part (b), we compute $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2$ and get

$$\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) &= \mathrm{Cov}((\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\boldsymbol{y}, (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{y}) \\
&= (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathrm{Var}[\boldsymbol{y}]\Big((\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\Big)^T \\
&= (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T(\sigma^2\mathbf{I})\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1} \\
&= \sigma^2(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1} = \mathbf{0}
\end{aligned}$$

Therefore, as $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ is jointly multivariate normal and $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) = \mathbf{0}$, $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ are independent.

For part (c), we note that the individual 95% confidence intervals for $\boldsymbol{a}^T\boldsymbol{\beta}_1$ and $\boldsymbol{b}^T\boldsymbol{\beta}_2$ are of the form

$$\begin{aligned}
\boldsymbol{a}^T\hat{\boldsymbol{\beta}}_1 \pm \mathrm{qt}(1-\alpha/2; n-p)\sqrt{\mathrm{MSE} \times \boldsymbol{a}^T(\mathbf{W}^T\mathbf{W})^{-1}\boldsymbol{a}} \\
\boldsymbol{b}^T\hat{\boldsymbol{\beta}}_2 \pm \mathrm{qt}(1-\alpha/2; n-p)\sqrt{\mathrm{MSE} \times \boldsymbol{b}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\boldsymbol{b}}
\end{aligned}$$

The two confidence intervals both depend on $\mathrm{MSE} = \|\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(n-p)$.

Problem 2: (10pts)

Let $\mathbf{W}$ be a $n \times p$ matrix and $\mathbf{X}$ be a $n \times q$ matrix and that $\mathcal{C}(\mathbf{W}) \subseteq \mathcal{C}(\mathbf{X})$. Denote by $\mathbf{P}_{\mathbf{X}}$ and $\mathbf{P}_{\mathbf{W}}$ the symmetric idempotent matrices projecting onto $\mathcal{C}(\mathbf{X})$ and $\mathcal{C}(\mathbf{W})$, respectively. Show that $\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{W}}$ is the symmetric orthogonal projection onto $\mathcal{C}((\mathbf{I} - \mathbf{P}_{\mathbf{W}})\mathbf{X})$.

You can do it by arguing as follows.

- First show that $\mathbf{P_X} - \mathbf{P_W}$ is idempotent. Hint: $\mathbf{P_X}\mathbf{P_W}\boldsymbol{z} = \mathbf{P_W}\boldsymbol{z}$ for all $\boldsymbol{z}$; in addition $\mathbf{P_W}\mathbf{P_X} = (\mathbf{P_X}\mathbf{P_W})^\top$.
- Next, show that for any $\boldsymbol{z}$, $(\mathbf{P_X} - \mathbf{P_W})\boldsymbol{z} \in \mathcal{C}((\mathbf{I} - \mathbf{P_W})\mathbf{X})$. Hint: Since $\mathcal{C}(\mathbf{X})$ and $\mathcal{N}(\mathbf{X}^\top)$ are orthogonal complements, any vector $\boldsymbol{z} \in \mathbb{R}^n$ can be written as $\boldsymbol{z} = \mathbf{X}\boldsymbol{v} + \boldsymbol{w}$ for some vectors $\boldsymbol{v}$ and some $\boldsymbol{w} \in \mathcal{N}(\mathbf{X}^\top)$; what is the relationship between $\mathcal{N}(\mathbf{W}^\top)$ and $\mathcal{N}(\mathbf{X}^\top)$?.
- Finally, show that if $\boldsymbol{z} \in \mathcal{C}((\mathbf{I} - \mathbf{P_W})\mathbf{X})$ then $(\mathbf{P_X} - \mathbf{P_W})\boldsymbol{z} = \boldsymbol{z}$.

Solution:

We first show that $\mathbf{P_X} - \mathbf{P_W}$ is idempotent. Indeed,

$$(\mathbf{P_X} - \mathbf{P_W})^2 = \mathbf{P_X} - \mathbf{P_X}\mathbf{P_W} - \mathbf{P_W}\mathbf{P_X} + \mathbf{P_W}.$$

Now, for any $\boldsymbol{z}$, $\mathbf{P_W}\boldsymbol{z} \in \mathcal{C}(\mathbf{W})$ and hence $\mathbf{P_W}\boldsymbol{z} \in \mathcal{C}(\mathbf{X})$. Thus $\mathbf{P_X}\mathbf{P_W}\boldsymbol{z} = \mathbf{P_W}\boldsymbol{z}$ for all $\boldsymbol{z}$ and hence $\mathbf{P_X}\mathbf{P_W} = \mathbf{P_W}$. Similarly, $\mathbf{P_W}\mathbf{P_X} = (\mathbf{P_X}\mathbf{P_W})^\top = \mathbf{P_W}^\top = \mathbf{P_W}$. We thus have $(\mathbf{P_X} - \mathbf{P_W})^2 = \mathbf{P_X} - \mathbf{P_W}$.

Next, for any $\boldsymbol{z} \in \mathbb{R}^n$, by the properties of orthogonal complements (and the fact that $\mathcal{C}(\mathbf{X})$ and $\mathcal{N}(\mathbf{X}^\top)$ are orthogonal complements), we have that $\boldsymbol{z} = \mathbf{X}\boldsymbol{v} + \boldsymbol{w}$ for some $\boldsymbol{v} \in \mathbb{R}^q$ and some $\boldsymbol{w} \in \mathcal{N}(\mathbf{X}^\top)$. We note also that $\mathcal{C}(\mathbf{W}) \subseteq \mathcal{C}(\mathbf{X})$ implies $\mathcal{N}(\mathbf{X}^\top) \subseteq \mathcal{N}(\mathbf{W}^\top)$. Then $(\mathbf{I} - \mathbf{P_X})\boldsymbol{w} = \boldsymbol{w}$ as $\mathbf{I} - \mathbf{P_X}$ is the orthogonal projection onto $\mathcal{N}(\mathbf{X}^\top)$. Therefore, $\mathbf{P_X}\boldsymbol{w} = \mathbf{P_X}(\mathbf{I} - \mathbf{P_X})\boldsymbol{w} = \mathbf{0}$. In addition, $(\mathbf{I} - \mathbf{P_W})\boldsymbol{w} = \boldsymbol{w}$ as $\mathbf{I} - \mathbf{P_W}$ is the orthogonal projection onto $\mathcal{N}(\mathbf{W}^\top) \supset \mathcal{N}(\mathbf{X}^\top)$ and hence $\mathbf{P_W}\boldsymbol{w} = \mathbf{0}$. Thus, for any $\boldsymbol{z}$

$$\begin{aligned}(\mathbf{P_X} - \mathbf{P_W})\boldsymbol{z} &= (\mathbf{P_X} - \mathbf{P_W})(\mathbf{X}\boldsymbol{v} + \boldsymbol{w}) \\ &= (\mathbf{P_X} - \mathbf{P_W})\mathbf{X}\boldsymbol{v} \\ &= \mathbf{X}\boldsymbol{v} - \mathbf{P_W}\mathbf{X}\boldsymbol{v} = (\mathbf{I} - \mathbf{P_W})\mathbf{X}\boldsymbol{v} \in \mathcal{C}((\mathbf{I} - \mathbf{P_W})\mathbf{X}).\end{aligned}$$

Finally, for any $\boldsymbol{z} \in \mathcal{C}((\mathbf{I} - \mathbf{P_W})\mathbf{X})$, we have $\boldsymbol{z} = (\mathbf{I} - \mathbf{P_W})\mathbf{X}\boldsymbol{v}$ for some $\boldsymbol{v}$ and hence

$$\begin{aligned}(\mathbf{P_X} - \mathbf{P_W})\boldsymbol{z} &= (\mathbf{P_X} - \mathbf{P_W})(\mathbf{I} - \mathbf{P_W})\mathbf{X}\boldsymbol{v} \\ &= (\mathbf{P_X} - \mathbf{P_W} - \mathbf{P_X}\mathbf{P_W} + \mathbf{P_W})\mathbf{X}\boldsymbol{v} \\ &= (\mathbf{P_X} - \mathbf{P_W})\mathbf{X}\boldsymbol{v} = (\mathbf{I} - \mathbf{P_W})\mathbf{X}\boldsymbol{v} = \boldsymbol{z}\end{aligned}$$

as desired.

Problem 3: (20pts)

The `kidiq.dta` dataset is available from the url `http://www.stat.columbia.edu/~gelman/arm/examples/child.iq/kidiq.dta` accompanying the book “Data Analysis using Regression and Multilevel/Hierarchical Models” by Gelman and

Hill. The dataset contains observations from a sample of 434 children. The variables include the child cognitive test scores at age 3 or 4, whether the mother finishes high school (coded as 1) or not (coded as 0), mother's IQ, age of mother at child's birth, and whether the mother work or not in the first three years of child's life. More specifically, the variable `mom.work` takes on the value

- `mom.work` = 1 if mother did not work in first three years of child's life
- `mom.work` = 2 if mother worked in second or third year of child's life
- `mom.work` = 3 if mother worked part-time in first year of child's life
- `mom.work` = 4 if mother worked full-time in first year of child's life

After downloading the `kidiq.dta` file you can read the data into `R` using the following snippet of code

```
library("foreign")
iq.data <- read.dta("kidiq.dta")
```

Using this dataset, answer the following questions.

(a) Perform a regression with `kid_score` as the response variable and the remaining variable except `mom_hs` as predictor variables.

(b) Provide a quick discussion regarding the coefficients for the predictor variables. What do they say ?

(c) Using the model in part [(a)], test the hypothesis that the predictor variables `mom_work` and `mom_age` is associated with the response variable. When do you recommend mothers should give birth ? What are your assumption for making this recommendation ?

(d) What happens when you add `mom_hs` as a predictor variable to the model in part (a) ? Have your conclusion about the timing of birth changed ?

(e) Using the model in part (d), perform some diagnostics, e.g., check the constant variance assumption, normality of errors. Look for outliers, influential points, and points with high leverage.

(f) Consider augmenting the model in part (d) with one whose predictor variables include interactions between say `mom.hs` and `mom_age` or interactions between say `mom.work` and `mom_age`. Write down the "formula" for the resulting model and discuss how it differs from the "formula" for the model in part (d). Test the hypothesis that the interaction term in the augmented model is not significant.

Solution:

```r
library("foreign")
iq.data <- read.dta("kidiq.dta")
```

```r
model1 <- lm(kid_score ~ mom_iq + factor(mom_work) + mom_age, iq.data)
summary(model1)
```

```
##
## Call:
## lm(formula = kid_score ~ mom_iq + factor(mom_work) + mom_age,
##     data = iq.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.158 -12.144   2.106  11.961  49.570
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       16.52456    9.30342   1.776   0.0764 .
## mom_iq             0.59045    0.05954   9.916   <2e-16 ***
## factor(mom_work)2  4.04858    2.79011   1.451   0.1475
## factor(mom_work)3  6.31045    3.25013   1.942   0.0528 .
## factor(mom_work)4  2.82054    2.45646   1.148   0.2515
## mom_age            0.35924    0.32956   1.090   0.2763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.23 on 428 degrees of freedom
## Multiple R-squared:  0.2113,Adjusted R-squared:  0.2021
## F-statistic: 22.94 on 5 and 428 DF,  p-value: < 2.2e-16
```

```r
model2 <- lm(kid_score ~ mom_iq, iq.data)
anova(model2, model1)
```

```
## Analysis of Variance Table
##
## Model 1: kid_score ~ mom_iq
## Model 2: kid_score ~ mom_iq + factor(mom_work) + mom_age
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    432 144137
## 2    428 142268  4    1869.4 1.406 0.2311
```

```r
model3 <- lm(kid_score ~ mom_hs + mom_iq + factor(mom_work) + mom_age, iq.data)
summary(model3)
```

```
##
## Call:
## lm(formula = kid_score ~ mom_hs + mom_iq + factor(mom_work) +
##     mom_age, data = iq.data)
##
## Residuals:
```

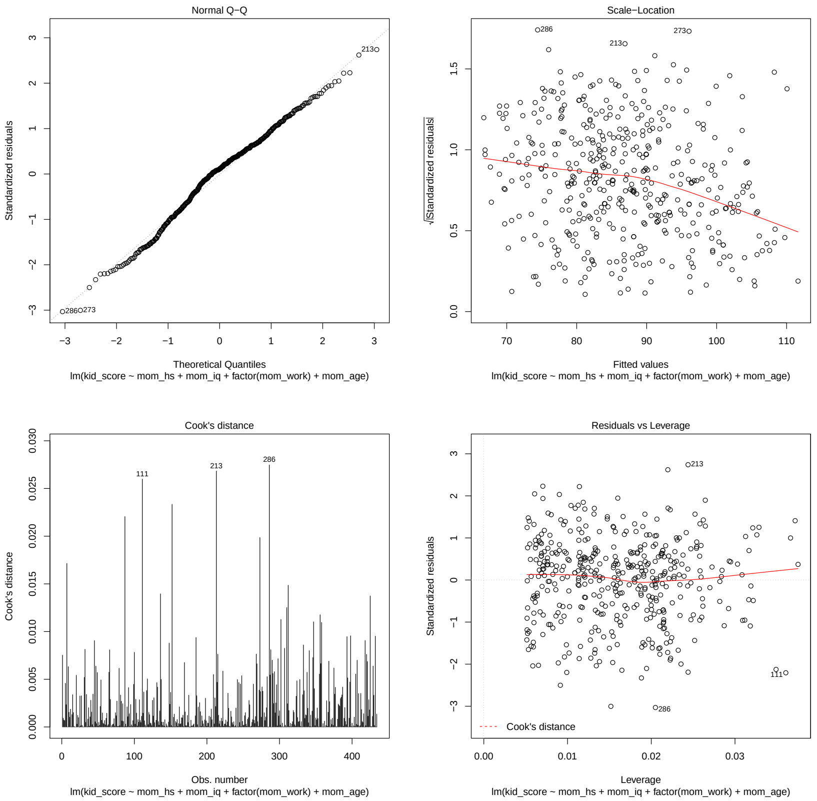
```
##     Min      1Q  Median      3Q     Max
## -54.414 -12.095   2.015  11.653  49.100
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       20.27273    9.39320   2.158   0.0315 *
## mom_hs             5.43466    2.32518   2.337   0.0199 *
## mom_iq             0.55288    0.06138   9.008   <2e-16 ***
## factor(mom_work)2  2.98266    2.81289   1.060   0.2896
## factor(mom_work)3  5.48824    3.25239   1.687   0.0922 .
## factor(mom_work)4  1.41929    2.51621   0.564   0.5730
## mom_age            0.21629    0.33351   0.649   0.5170
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.14 on 427 degrees of freedom
## Multiple R-squared:  0.2213,Adjusted R-squared:  0.2103
## F-statistic: 20.22 on 6 and 427 DF,  p-value: < 2.2e-16
```

```
table(iq.data$mom_hs, iq.data$mom_work)
```

```
##
##       1   2   3   4
##   0  33  23  11  26
##   1  44  73  45 179
```

```
plot(model3, which = c(2:5))
model4 <- lm(kid_score ~ mom_iq + mom_hs * mom_age + factor(mom_work), iq.data)
anova(model3, model4)
```

```
## Analysis of Variance Table
##
## Model 1: kid_score ~ mom_hs + mom_iq + factor(mom_work) + mom_age
## Model 2: kid_score ~ mom_iq + mom_hs * mom_age + factor(mom_work)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    427 140471
## 2    426 139247  1    1223.3 3.7425 0.05371 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



The model `model1` is the regression model for part (a). The coefficients for `model1` indicate that `mom_iq` is the only statistically significant predictor variable, assuming that the error terms are independent and have constant variance; also, an increase in `mom_iq` is associated with increase in the response variable.

For part (c), the call to `anova(model1, model2)` indicates that, in the presence of `mom_iq`, the various levels of the *categorical* variable `mom_work` and the variable `mom_age` are not particularly useful in predicting the response variable. Nevertheless, suppose that we can make the claim that since the coefficient of `mom_age` is positive, that all else is equal, giving birth at a later age is "preferable". This recommendation, however, depends on a lot of strong assumptions that is very hard to justify. The most obvious criticism is that as the estimated coefficient of `mom_age` is only 0.36, delaying birth by say 5 years will only be associated with say a potential increase of 2 points on `kid_score`. As the estimated $\sqrt{\mathrm{MSE}} \approx 18.24$, this "increase" is minimal compared to the estimated variability.

For part (d), we see that adding `mom_hs` yield a model that is not much different from the original model in part (a). While the coefficient of `mom_hs` can be claimed to be statistically significant in the presence of other variables, the effect on the least square fit is minimal, as the change in R^2 can attest. Furthermore, there appear to be some hint of multicollinearity between the predictor variable `mom_hs` and the dummy variable corresponding to `mom_work` $= 4$. A quick check reveals that for a majority of observations, `mom_work` $= 4$ when `mom_hs` $= 1$ and vice versa.

For part (e), the various residual plots suggest that there are no sign of non-constant variance, no sign of outliers, and no sign of influential points. There are a few observations with leverage somewhat bigger than the conventional threshold of $2p/n$ where p is the number of predictor variables (including the intercept), and n is the number of observations. Nevertheless, for 434 data points, it is usually hard to claim that any small collection of data point is really influential (unless they have “extreme” leverage). We can try removing some of the data points with largest leverage and refit the model; it turns out that the coefficient estimate and the estimated mean square error only change slightly. The assumption that the error terms are normally distributed also seems plausible.

The interaction term between `mom_hs` and `mom_age` is incorporated into `model4`. This interaction term does not appear to be significant in the presence of the remaining terms. By incorporating the interaction term, we allow for the fact that the *coefficient* for `mom_age` in `model4` can depend on whether `mom_hs` $= 1$ or not.

Problem 4: (20pts)

The link `http://www.amstat.org/publications/jse/v16n3/kuiper.xls` is a dataset collected from the Kelly Blue Book for several hundred 2005 used GM cars. Do something with this data. This is meant to be an open-ended question. For some ideas of the kind of analysis one can attempt, see the article `http://www.amstat.org/publications/jse/v16n3/datasets.kuiper.html`.

Solution:

Roughly anything is fine. I just hope it is not (to paraphrase H. L. Mencken)

```
I was at the job of working on it for days and days, endlessly daunted
and halted by its laborious dullness, its flatulent fatuity, its
almost fabulous inconsequentiality.
```

The main ideas to grok from this problem are (1) working with categorical variables, (2) using regression diagnostics to inform subsequent modeling, (3) there are many models that can be considered and that many of these models are “equivalent”, at least with respect to the mean square error criterion, and (4) identify potential signs of multi-collinearity.