Assignment 6 Answer Sketch

431 staff

due 2017-12-04 at noon

Contents

Grading: Questions 2, 3, 4 and 5 are graded, for a total possible 70 points.

R Setup

```r
knitr::opts_chunk$set(comment=NA)
options(width = 75)

library(forcats); library(pander); library(car)
library(gridExtra); library(tidyverse)

plasma <- read.csv("hw6plasma.csv") %>% tbl_df

plasma$sex <- fct_recode(as.character(plasma$sex), "Female" = "2", "Male" = "1")

plasma$smoking <- fct_recode(as.character(plasma$smoking),
                "never" = "1", "former" = "2", "current" = "3")
```

Question 1 (optional - graded as Yes or No)

Find a Wikipedia page with a table that interests you. Scrape the data from the table and clean it up, converting character strings to numbers, dropping unneeded variables, and so forth, leading to a tidy data set in R. Create a statistical graph from these data. Write a caption for your figure that interprets your findings. Be sure to provide the link to the original Wikipedia table.

We didn't write a sketch for this question.

Students got a Yes if [a] they addressed the specified tasks in an easy-to-follow manner, [b] their Wikipedia page was linked properly, [c] they wrote an appropriate caption and [d] they produced a reasonable graph [e] from a tidy data set. If most, but not quite all of those things were true, we still marked Yes, but a substantial effort towards each piece was required. Otherwise, it's a No.

Question 2 (30 points)

Find an example of a visualization designed to support a linear regression analysis in a published work (online or not) for which you can find the complete sourcing information, and which was built no earlier than January 1, 2014. Provide the complete reference and a copy of the image itself (including any captions or titles) and surrounding material for the visualization, and provide a brief essay (likely to run 150-250 words) which accomplishes each of the following tasks:

- *Describe the linear regression model behind the visualization. Explain its context and why it is important. Specify the research question that this regression model answers.*
- *Describe the visualization and explain what you believe it is trying to do. Specify why it is or is not effective, in your view.*
- *Provide your best suggestion as to how the visualization might be improved, and explain why your change would be an improvement.*

We didn't write a sketch for this question.

Question 3 (15 points)

*In questions 3-6, you will build and present an appropriate model for plasma retinol levels. Use only the **275 observations** where **`holdout`** is 0 for Questions 3-9.*

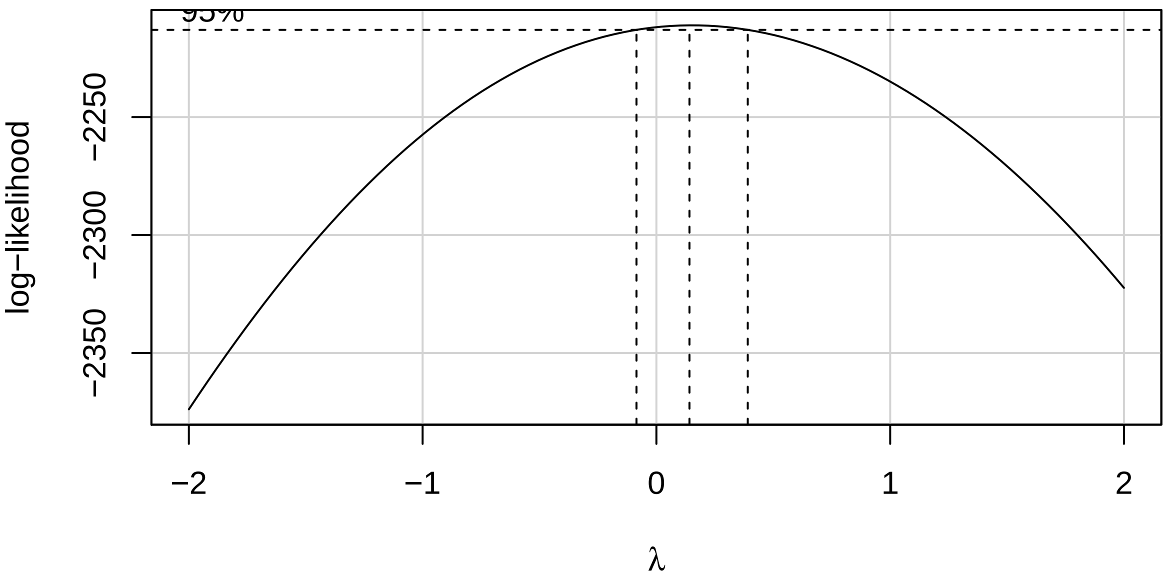
```
plasma_dev <-
    plasma %>%
    filter(holdout == 0)
```

a. *Specify whether a transformation of the outcome is necessary, and how you know. If you need a transformation, specify a wise choice.*

I had intended that you would first build the necessary regression model and use the Box-Cox procedure to identify potential outcome transformations…

```
ret.m1 <- lm(retplasma ~ age + sex + factor(smoking) + bmi
              + factor(vitamin) + calories + fat + fiber
              + alcohol + cholesterol + retdiet, data = plasma_dev)

boxCox(ret.m1)
```



```
powerTransform(ret.m1)
```

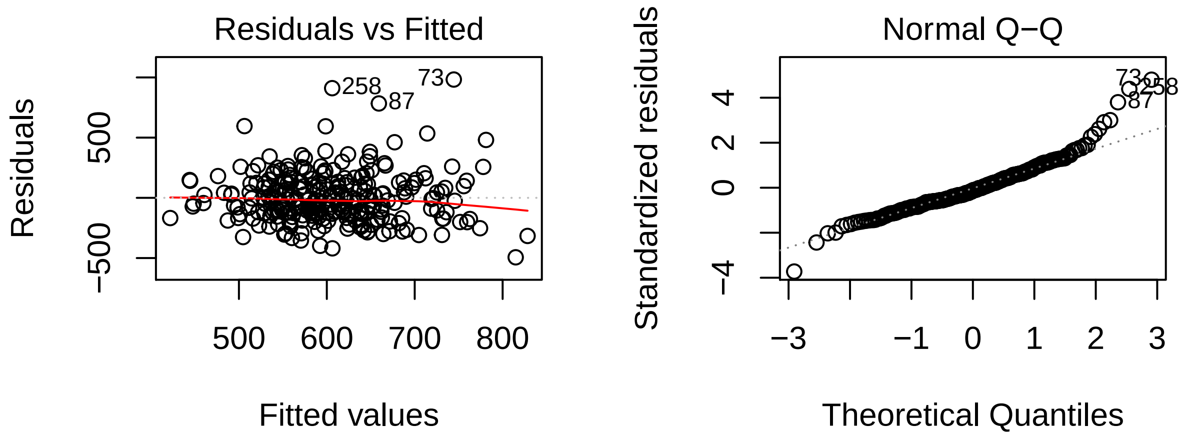
```
Estimated transformation parameters
       Y1
0.1518265
```

This approach suggests using the logarithm of `retplasma` as our outcome. Here is that model.

```
ret.m2 <- lm(log(retplasma) ~ age + sex + factor(smoking) + bmi
              + factor(vitamin) + calories + fat + fiber
              + alcohol + cholesterol + retdiet, data = plasma_dev)
```

Now, we can compare the residual plots generated by these two models.

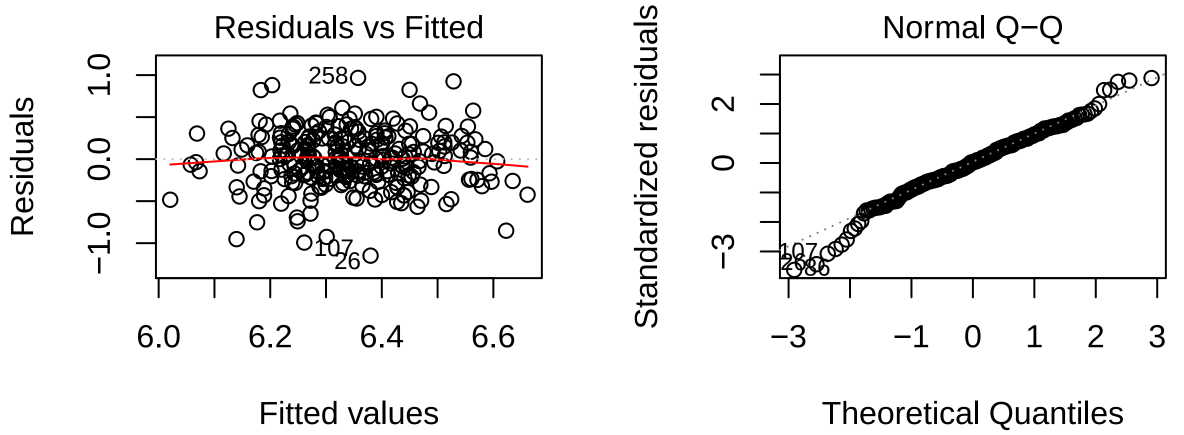
```r
par(mfrow = c(1,2))
plot(ret.m1, which = 1:2)
```



```r
par(mfrow = c(1,1))
```

and then, after transforming the outcome (`retplasma`) using the logarithm, we have the following residual plots.

```r
par(mfrow = c(1,2))
plot(ret.m2, which = 1:2)
```



```r
par(mfrow = c(1,1))
```

In my mind, the logarithm is better, in that we have a better balance above and below the zero line on the left plot (residuals vs. fitted plot) and a somewhat better Normal Q-Q plot, especially at the top end of the distribution. But you could probably go either way here. I'll show the remaining results with the log transformation for the outcome.

b. *Select predictors from the demographic, behavioral and relevant dietary factors described in the data (i.e. not including* ***id****,* ***betadiet****,* ***betaplasma*** *or* ***holdout****.)*

We'll start with our "kitchen sink" model.

```
ret.m2 <- lm(log(retplasma) ~ age + sex + factor(smoking) + bmi
              + factor(vitamin) + calories + fat + fiber
              + alcohol + cholesterol + retdiet, data = plasma_dev)
summary(ret.m2)
```

```

Call:
lm(formula = log(retplasma) ~ age + sex + factor(smoking) + bmi +
    factor(vitamin) + calories + fat + fiber + alcohol + cholesterol +
    retdiet, data = plasma_dev)

Residuals:
     Min       1Q   Median       3Q      Max
-1.14873 -0.19927  0.00879  0.22585  0.96703

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.231e+00  1.736e-01  35.892  < 2e-16 ***
age                     5.722e-03  1.596e-03   3.585 0.000403 ***
sexFemale              -7.968e-02  6.720e-02  -1.186 0.236810
factor(smoking)former   1.044e-01  4.501e-02   2.319 0.021193 *
factor(smoking)current -2.691e-02  7.095e-02  -0.379 0.704855
bmi                    -1.181e-03  3.470e-03  -0.340 0.733847
factor(vitamin)2        4.985e-02  5.301e-02   0.940 0.347867
factor(vitamin)3        6.700e-03  4.988e-02   0.134 0.893255
calories                1.444e-04  1.085e-04   1.330 0.184603
fat                    -2.183e-03  1.721e-03  -1.268 0.205890
fiber                  -1.113e-02  5.627e-03  -1.979 0.048883 *
alcohol                -3.396e-03  2.560e-03  -1.326 0.185860
cholesterol            -2.442e-04  2.287e-04  -1.068 0.286579
retdiet                -5.174e-06  3.813e-05  -0.136 0.892167
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3389 on 261 degrees of freedom
Multiple R-squared:  0.1133,	Adjusted R-squared:  0.06919
F-statistic: 2.567 on 13 and 261 DF,  p-value: 0.002361
```

Next, we'll use a stepwise algorithm to identify a set of predictors worthy of further study.

```
ret.m3 <- step(ret.m2)
```

```
Start:  AIC=-581.58
log(retplasma) ~ age + sex + factor(smoking) + bmi + factor(vitamin) +
    calories + fat + fiber + alcohol + cholesterol + retdiet

                  Df Sum of Sq    RSS     AIC
- factor(vitamin)  2   0.11211 30.080 -584.55
- retdiet          1   0.00211 29.970 -583.56
- bmi              1   0.01330 29.981 -583.45
- cholesterol      1   0.13093 30.099 -582.38
- sex              1   0.16143 30.129 -582.10
- fat              1   0.18464 30.153 -581.89
- alcohol          1   0.20201 30.170 -581.73
- calories         1   0.20318 30.171 -581.72
<none>                         29.968 -581.58
- fiber            1   0.44962 30.418 -579.48
- factor(smoking)  2   0.77996 30.748 -578.51
- age              1   1.47530 31.443 -570.36

Step:  AIC=-584.55
log(retplasma) ~ age + sex + factor(smoking) + bmi + calories +
    fat + fiber + alcohol + cholesterol + retdiet

                  Df Sum of Sq    RSS     AIC
- retdiet          1   0.00275 30.083 -586.52
- bmi              1   0.00771 30.088 -586.48
- cholesterol      1   0.12197 30.202 -585.44
- sex              1   0.15086 30.231 -585.17
- fat              1   0.20738 30.288 -584.66
- alcohol          1   0.20942 30.290 -584.64
- calories         1   0.21694 30.297 -584.57
<none>                         30.080 -584.55
- fiber            1   0.45233 30.532 -582.44
- factor(smoking)  2   0.79342 30.873 -581.39
- age              1   1.40071 31.481 -574.03

Step:  AIC=-586.52
log(retplasma) ~ age + sex + factor(smoking) + bmi + calories +
    fat + fiber + alcohol + cholesterol

                  Df Sum of Sq    RSS     AIC
- bmi              1   0.00764 30.090 -588.45
- cholesterol      1   0.13903 30.222 -587.26
- sex              1   0.15318 30.236 -587.13
- alcohol          1   0.20707 30.290 -586.64
- fat              1   0.20806 30.291 -586.63
- calories         1   0.21485 30.298 -586.57
<none>                         30.083 -586.52
- fiber            1   0.45761 30.541 -584.37
- factor(smoking)  2   0.79743 30.880 -583.33
- age              1   1.40038 31.483 -576.01

Step:  AIC=-588.45
log(retplasma) ~ age + sex + factor(smoking) + calories + fat +
    fiber + alcohol + cholesterol

                  Df Sum of Sq    RSS     AIC
- cholesterol      1   0.14569 30.236 -589.13
- sex              1   0.15355 30.244 -589.05
- alcohol          1   0.20130 30.292 -588.62
- fat              1   0.20636 30.297 -588.58
- calories         1   0.21135 30.302 -588.53
<none>                         30.090 -588.45
- fiber            1   0.45001 30.541 -586.37
- factor(smoking)  2   0.80281 30.893 -585.21
- age              1   1.39562 31.486 -577.99

Step:  AIC=-589.13
log(retplasma) ~ age + sex + factor(smoking) + calories + fat +
    fiber + alcohol

                  Df Sum of Sq    RSS     AIC
- sex              1   0.11222 30.348 -590.11
- calories         1   0.16107 30.397 -589.67
- alcohol          1   0.16969 30.406 -589.59
<none>                         30.236 -589.13
- fat              1   0.29433 30.531 -588.46
- fiber            1   0.38592 30.622 -587.64
- factor(smoking)  2   0.82588 31.062 -585.72
- age              1   1.40867 31.645 -578.60

Step:  AIC=-590.11
log(retplasma) ~ age + factor(smoking) + calories + fat + fiber +
    alcohol

                  Df Sum of Sq    RSS     AIC
- alcohol          1   0.14701 30.495 -590.78
- calories         1   0.16444 30.513 -590.62
<none>                         30.348 -590.11
- fat              1   0.27247 30.621 -589.65
- fiber            1   0.39233 30.741 -588.58
- factor(smoking)  2   0.87052 31.219 -586.33
- age              1   1.83930 32.188 -575.93

Step:  AIC=-590.78
log(retplasma) ~ age + factor(smoking) + calories + fat + fiber

                  Df Sum of Sq    RSS     AIC
- calories         1   0.03169 30.527 -592.49
- fat              1   0.12929 30.625 -591.62
<none>                         30.495 -590.78
- fiber            1   0.25007 30.745 -590.53
- factor(smoking)  2   0.85569 31.351 -587.17
- age              1   1.69230 32.188 -577.93

Step:  AIC=-592.49
log(retplasma) ~ age + factor(smoking) + fat + fiber

                  Df Sum of Sq    RSS     AIC
- fat              1   0.16510 30.692 -593.01
<none>                         30.527 -592.49
- fiber            1   0.22298 30.750 -592.49
- factor(smoking)  2   0.84408 31.371 -588.99
- age              1   1.66227 32.189 -579.91

Step:  AIC=-593.01
log(retplasma) ~ age + factor(smoking) + fiber
```

```
                  Df Sum of Sq    RSS     AIC
<none>                         30.692 -593.01
- fiber            1   0.41336 31.106 -591.33
- factor(smoking)  2   0.82329 31.515 -589.73
- age              1   1.87911 32.571 -578.67
```

```
pander(summary(ret.m3), caption = "Summary of Stepwise model", digits = 3)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.12	0.0914	66.9	4.61e-170
age	0.00569	0.0014	4.07	6.29e-05
factor(smoking)former	0.106	0.0438	2.42	0.016
factor(smoking)current	-0.0291	0.0667	-0.436	0.663
fiber	-0.00742	0.00389	-1.91	0.0576

Table 2: Summary of Stepwise model

Observations	Residual Std. Error	R^2	Adjusted R^2
275	0.3372	0.09192	0.07847

c. *Motivate your choice of predictors, including an assessment of the impact of collinearity, with appropriate accounting for it in your final choice of model. Specify and demonstrate the impact of your model selection algorithm.*

I used a stepwise algorithm (backwards elimination) to obtain the models above. In each case, I can identify possible collinearity issues using the variance inflation factor.

```
car::vif(ret.m3)
```

```
                    GVIF Df GVIF^(1/(2*Df))
age             1.022460  1        1.011167
factor(smoking) 1.046595  2        1.011451
fiber           1.025001  1        1.012424
```

Note that this version of the `vif` function, obtained through the `car` library, calculates a generalized variance inflation factor. We see no signs of substantial collinearity here.

Question 4 (15 points, 5 points for each part)

a. *Summarize the findings in a clear presentation of your final model, including a short recap of the steps you took to produce it.*

Here, I was looking for a detailed statement of the model, indicating its overall significance, the R-squared value and the directions of the coefficients in the model. In the holdout sample, we might conclude that the combination of age, smoking status and fiber, together explained just over 9% of the variation in plasma retinol, and that this model had statistically significant predictive value. We would then indicate that retinol was associated with increasing age, being a former smoker (as opposed to a current or never smoker), while it is also associated with decreasing fiber consumption. We might even have standardized the coefficients to look at the relative size of each predictor's impact on the predictions.

b. *Demonstrate the utility of the final model, including summaries based on R^2 and significance testing.*

The R-squared for this model is 0.092 meaning that the model accounts for 9.2% of the variation in plasma concentration of retinol.

The adjusted R-squared for the model is 0.078 which indicates no severe issues with overfitting.

In terms of significance testing, we have the following t test results:

```
pander(summary(ret.m3))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.115	0.09144	66.88	4.611e-170
age	0.005693	0.0014	4.066	6.285e-05
factor(smoking)former	0.1062	0.04381	2.424	0.01602
factor(smoking)current	-0.02909	0.06669	-0.4362	0.6631
fiber	-0.007421	0.003892	-1.907	0.05759

Table 4: Fitting linear model: log(retplasma) ~ age + factor(smoking) + fiber

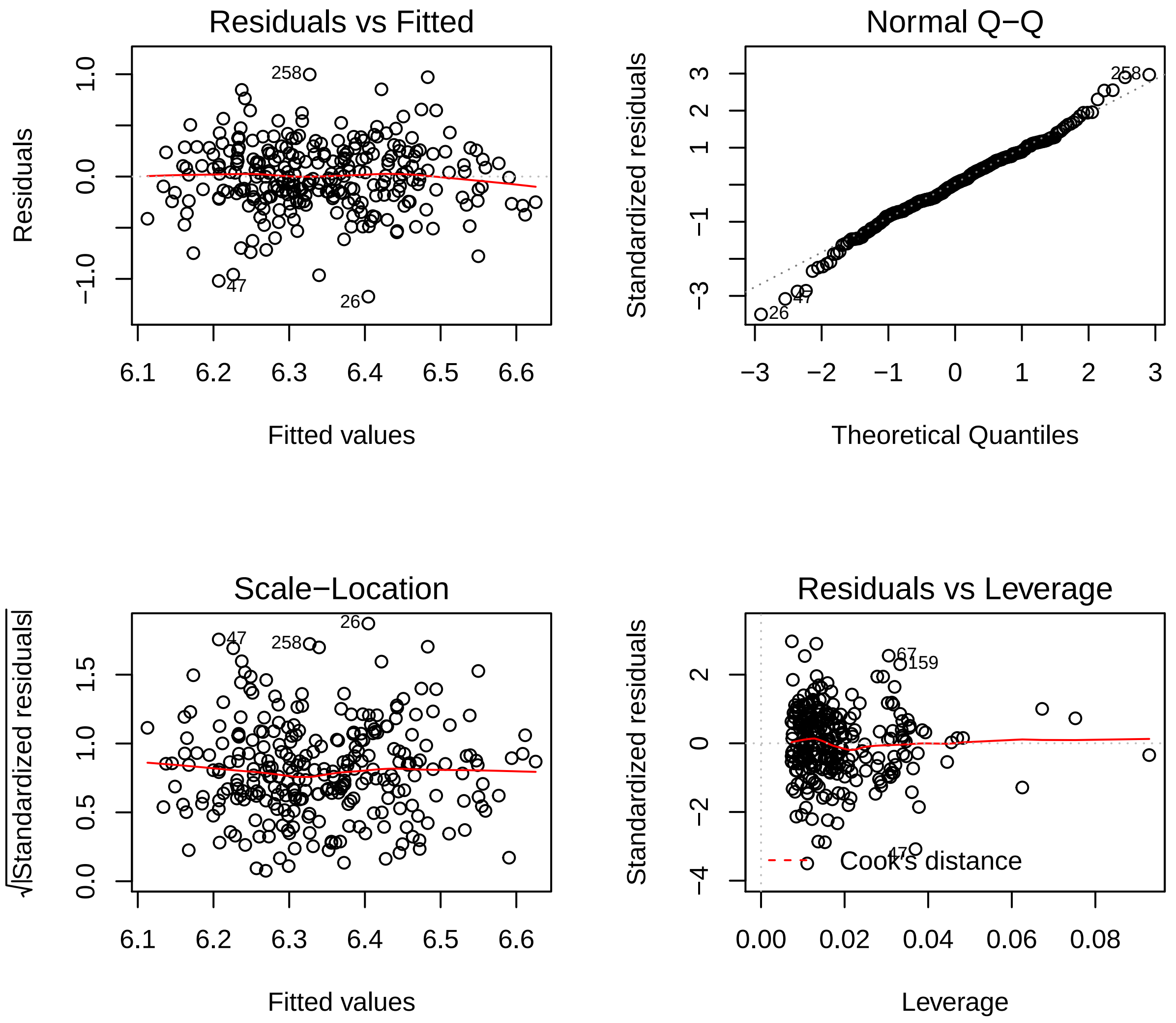
Observations	Residual Std. Error	R^2	Adjusted R^2
275	0.3372	0.09192	0.07847

c. *Demonstrate that your final model passes required checks of assumptions.*

A set of diagnostic model plots will inform the answer to this question:

```
par(mfrow = c(2, 2))

plot(ret.m3)
```



```r
par(mfrow = c(1,1))
```

There are no signs of very serious trouble here with the assumptions of linearity, normality (maybe a few outliers on the low end of the residual distribution), homoscedasticity, and independence isn't really an issue here.

Question 5 (10 points)

In a single sentence, outline the key findings for plasma retinol specified by your model.

This would be a shorter description - a subset of the previous answer would do - something like: Retinol was associated with increasing age and being a former smoker, while it is also associated with decreasing fiber consumption.

Question 6 (optional - graded as Yes or No)

At this point, we will return to working with the whole set of 300 observations. Validate your choice of model for plasma retinol level by using your final choice developed in questions 3-9 to predict data for the 25 cases that you have withheld from the data, comparing your final model to these two other models:

- *A model using `age` and `sex` alone.*
- *A model that uses the entire set (kitchen sink) of possible predictors, i.e. everything but `id`, `betadiet` and `betaplasma`.*

Which model looks best in your comparison? Justify your response.

Our goal now is to compare the predictions made by our final choice of model for plasma retinol established in Question 5 above to those of a model with age and sex only, and to the model with all available predictors using our holdout sample of 25 observations, and using both mean absolute prediction error and mean squared prediction error to develop our conclusions.

Our three models are

```
model1 <- lm(log(retplasma) ~ age + sex, data=plasma_dev)

model2 <- ret.m3

model3 <- lm(log(retplasma) ~ age + sex + factor(smoking) + bmi
              + factor(vitamin) + calories + fat + fiber
              + alcohol + cholesterol + retdiet, data=plasma_dev)

anova(model1, model2, model3)
```

```
Analysis of Variance Table

Model 1: log(retplasma) ~ age + sex
Model 2: log(retplasma) ~ age + factor(smoking) + fiber
Model 3: log(retplasma) ~ age + sex + factor(smoking) + bmi + factor(vitamin) +
    calories + fat + fiber + alcohol + cholesterol + retdiet
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    272 31.780
2    270 30.692  2   1.08813 4.7384 0.009519 **
3    261 29.968  9   0.72422 0.7008 0.707968
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA suggests that dropping from model 3 to model 2 here is not causing a statistically significant lack of fit, but that moving from model 2 down to model 1 might be (we can't be sure because Model 1 is not a subset of Model 2.

Now, we'll predict the values of `retplasma` in the holdout sample using each of the three models in turn, then compare the errors made, and calculate the absolute errors and squared errors for each prediction. Since our models predict the log of our outcome, we will exponentiate the predictions to get back on the original `retplasma` scale.

```
plasma_test <- plasma %>% filter(holdout == 1)

model1.pre <- exp(predict(model1, newdata=plasma_test))
model2.pre <- exp(predict(model2, newdata=plasma_test))
model3.pre <- exp(predict(model3, newdata=plasma_test))

model1.err <- plasma_test$retplasma - model1.pre
```

```
model2.err <- plasma_test$retplasma - model2.pre
model3.err <- plasma_test$retplasma - model3.pre

model1.abserr <- abs(model1.err)
model2.abserr <- abs(model2.err)
model3.abserr <- abs(model3.err)

model1.sqerr <- model1.err^2
model2.sqerr <- model2.err^2
model3.sqerr <- model3.err^2
```

Now, we summarize the prediction errors in the holdout sample for predicting retinol plasma concentrations.

Model	MAPE	MSPE
1 (age + sex)	119.6	23733
2 (our model)	122.6	27041
3 (kitchen sink)	122.1	26005

Looking at the mean absolute prediction error (MAPE) and the mean squared prediction error (MSPE), it appears that the `age + sex` model is superior to the competitors in this test sample.