HW3

Your Name Goes Here

Details

Due Date

This assignment is due in two phases:

- You must commit and push preliminary answers to problems 1 and 2 by 9:30 AM on Friday, Oct 2. I will be checking to make sure you have done this, and some of the completion points are for this preliminary commit. I'll be looking for thorough and complete answers to these questions.
- The full assignment is due at 5:00 PM Saturday Oct 3. You must commit and push final answers to all problems by this time. It's ok if you change your preliminary answers from your first commit.

The reason I'm doing this is that the problem set is somewhat long, so I want to make sure you get started early and make substantial progress before the due date so that you can ask questions in office hours. But I also recognize that we have a lot going on this week and it's a busy time of the semester.

Grading

20% of your grade on this assignment is for completion. A quick pass will be made to ensure that you've made a reasonable attempt at all problems.

Some of the problems will be graded more carefully for correctness. In grading these problems, an emphasis will be placed on full explanations of your thought process. You usually won't need to write more than a few sentences for any given problem, but you should write complete sentences! Understanding and explaining the reasons behind your decisions is more important than making the "correct" decision.

Solutions to all problems will be provided.

Collaboration

You are allowed to work with others on this assignment, but you must complete and submit your own write up. You should not copy large blocks of code or written text from another student.

Sources

You may refer to class notes, our textbook, Wikipedia, etc.. All sources you refer to must be cited in the space I have provided at the end of this problem set.

In particular, you may find the following resources to be valuable:

- Courses assigned on DataCamp
- Example R code from class
- Cheat sheets and resources linked from [http://www.evanlray.com/stat340_f2019/resources.html]

Load Packages

The following R code loads packages needed in this assignment.

```
library(readr)
library(dplyr)
library(ggplot2)
library(GGally)
library(gridExtra)
```

Note: The conceptual problems come after the applied questions in this assignment. It will be helpful to you to have done the applied problem before you do the conceptual problems.

Applied Problems

Problem 1: Smoking and Pulmonary Function

We will analyze data that originally appeared in the following two studies investigating health effects of smoking.

Tager, I., Weiss, S., Munoz, A., Rosner, B., and Speizer, F. (1983), "Longitudinal Study of the Effects of Maternal Smoking on Pulmonary Function," New England Journal of Medicine, 309(12), 699-703.

Tager, I., Weiss, S., Rosner, B., and Speizer, F. (1979), "Effect of Parental Cigarette Smoking on the Pulmonary Function of Children," American Journal of Epidemiology, 110(1), 15-26.

The data are from a sample of 654 youths, aged 3 to 19, in the area of East Boston. We have observations on the following variables:

- age: the subject's age in years
- fev: forced expiratory volume (FEV), the amount of air an individual can exhale in the first second of a forceful breath, in units of liters.
- height: the subject's height in inches
- sex: the subject's self-reported sex
- smoke: "smoker" if the subject reports smoking cigarettes regularly, "nonsmoker" otherwise

Our goal is to understand the relationship between an individual's smoking status and the health of their lungs, as measured by FEV. Roughly, a larger value of FEV indicates greater lung capacity and healthier lungs. FEV will be our response variable in this analysis.

The following R code reads the data in.

```
fev_data <- read_table("http://www.evanlray.com/data/jse/fev/fev.dat.txt", col_names = FALSE)
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   X2 = col_double(),
##   X3 = col_double(),
##   X4 = col_double(),
##   X5 = col_double()
## )
```

```
names(fev_data) <- c("age", "fev", "height", "sex", "smoke")

fev_data <- fev_data %>%
  mutate(
    sex = factor(sex),
    smoke = factor(smoke)
  )
levels(fev_data$sex) <- c("female", "male")
levels(fev_data$smoke) <- c("nonsmoker", "smoker")

head(fev_data)
```

```
## # A tibble: 6 x 5
##     age   fev height sex    smoke
##   <dbl> <dbl>  <dbl> <fct>  <fct>
## 1     9  1.71   57   female nonsmoker
## 2     8  1.72   67.5 female nonsmoker
## 3     7  1.72   54.5 female nonsmoker
## 4     9  1.56   53   male   nonsmoker
## 5     9  1.90   57   male   nonsmoker
## 6     8  2.34   61   female nonsmoker
```

(a) Univariate analysis

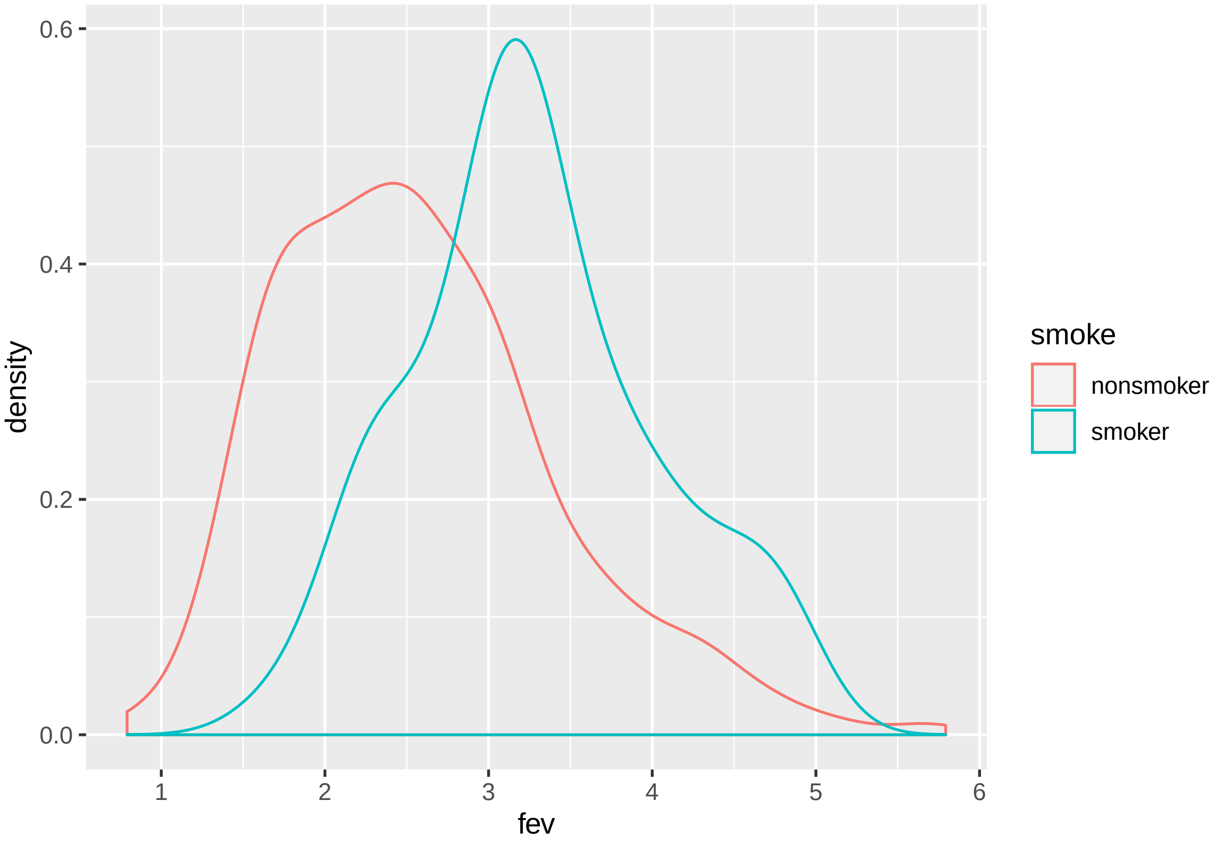
First, let's conduct an analysis with only the explanatory variable of primary interest, smoking status.

i. Plot the data

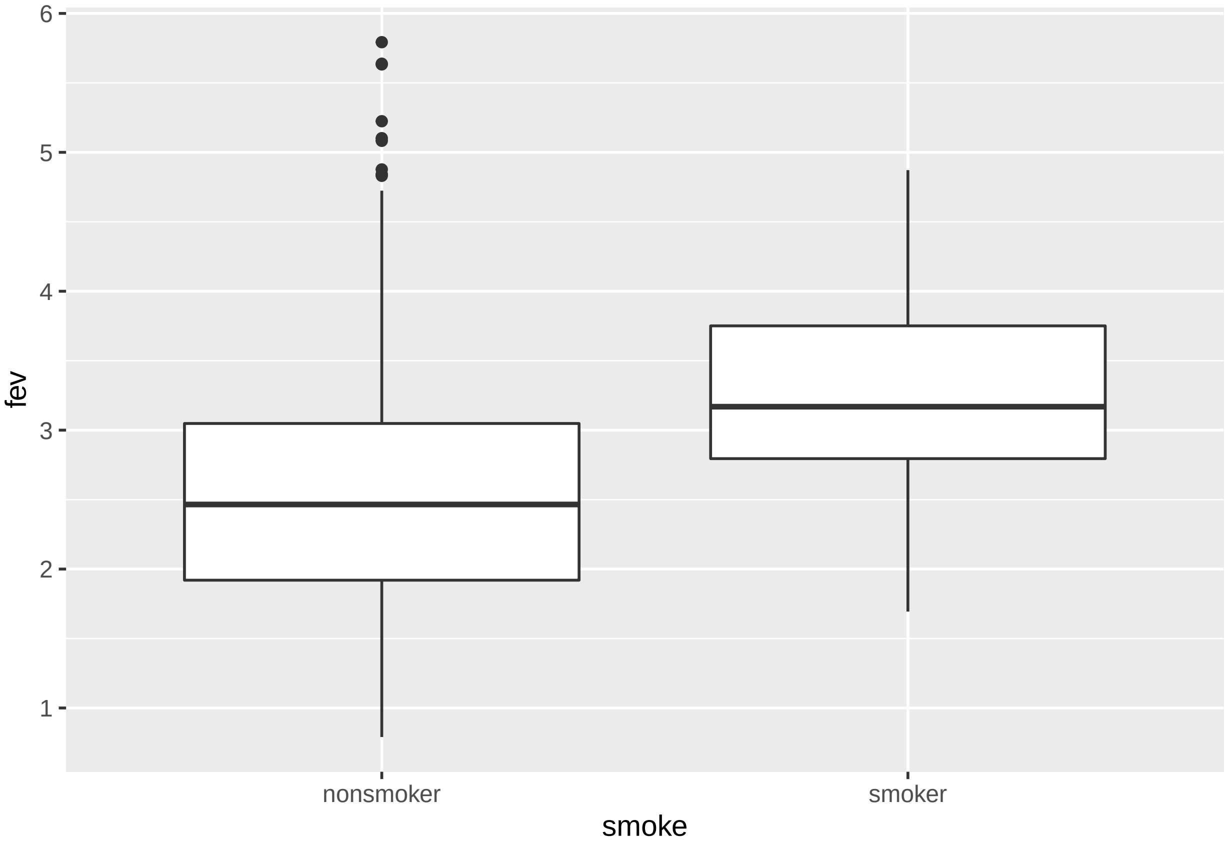
Make a plot comparing the observed FEV values in the smoker group and the nonsmoker group.

You have many options for what plot to make; here are a couple.

```
ggplot(data = fev_data, mapping = aes(x = fev, color = smoke)) +
  geom_density()
```



```
ggplot(data = fev_data, mapping = aes(x = smoke, y = fev)) +
  geom_boxplot()
```



ii. Fit a model

Also fit an ANOVA-type model, with `fev` as the response variable and `smoke` as the explanatory variable.

```
fit <- lm(fev ~ smoke, data = fev_data)
summary(fit)
```

```
##
## Call:
## lm(formula = fev ~ smoke, data = fev_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7751 -0.6339 -0.1021  0.4804  3.2269
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.56614    0.03466  74.037  < 2e-16 ***
## smokesmoker  0.71072    0.10994   6.464 1.99e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8412 on 652 degrees of freedom
## Multiple R-squared:  0.06023,    Adjusted R-squared:  0.05879
## F-statistic: 41.79 on 1 and 652 DF,  p-value: 1.993e-10
```

iii. Explain your findings

Discuss what your analysis has to say about the relationship between smoking and FEV. This discussion should include interpretions of the coefficient estimates and confidence intervals for terms in your final model fit. In writing this up, target a public health researcher who is not an expert in statistics.

Both the plot and the model fit show that in this data set, `fev` tends to be larger for children who smoke than it is for children who do not smoke. We see this from the density plot since the density estimate for smokers is shifted to the right of the density estimate for nonsmokers. We see this from the model fit since the estimated coefficient labeled `smokesmoker` is positive. This coefficient describes the difference in predicted FEV between smokers and nonsmokers; the positive coefficient value indicates that the estimated mean FEV is higher for smokers than it is for nonsmokers. Moreover, a t test of the hypotheses $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ has p-value 1.99e-10, indicating extremely strong evidence against the null hypothesis that mean FEV is the same for both groups.

iv. Reflect

Does the conclusion you've drawn from this analysis seem "correct" in the context of the data?

This difference does not match with what I might expect to see. A higher FEV indicates healthier lungs. I would not expect smokers to have healthier lungs than nonsmokers.

(b) Multivariate analysis

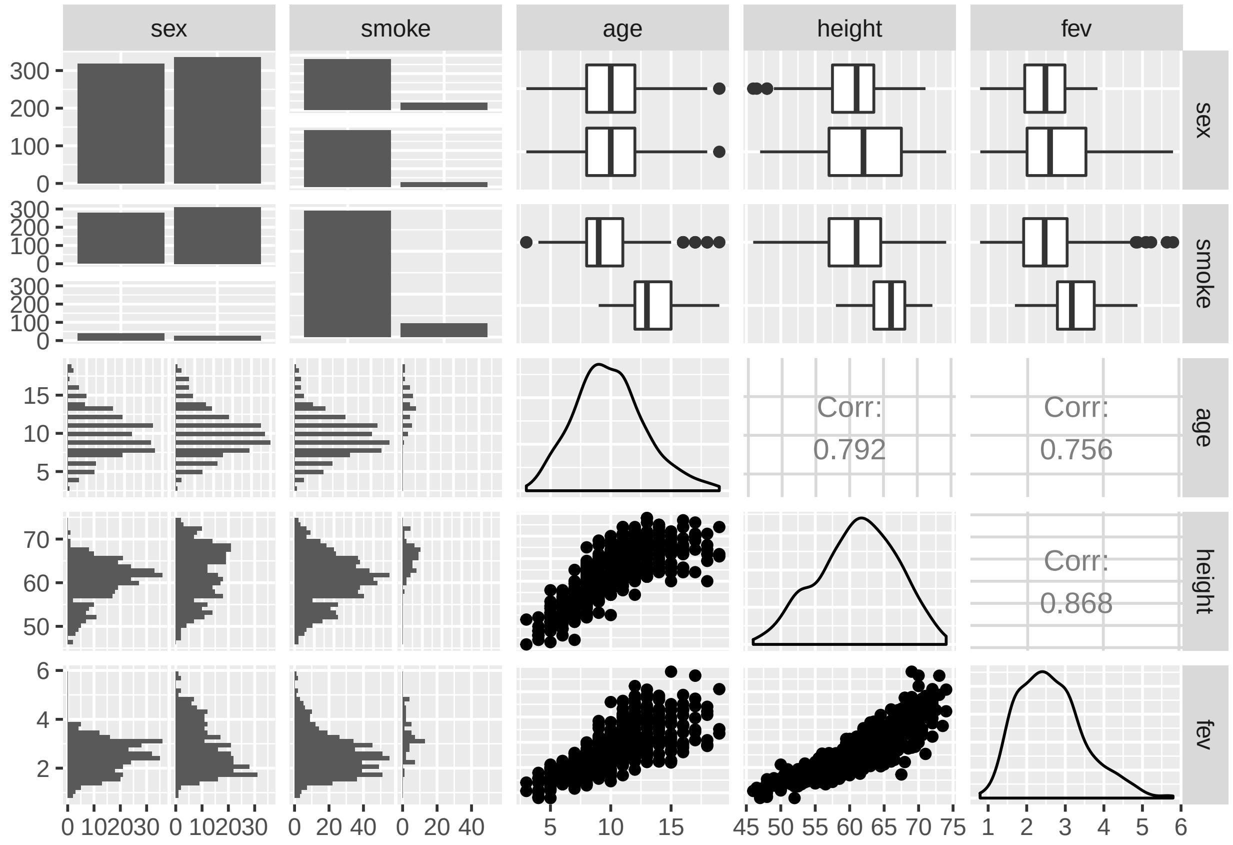
These data come from an observational study, so in studying the relationship between smoking status and FEV we should control for the effects of any other variables that are also associated with FEV. The big question is what other variables need to be included in the model, and exactly how they should enter (do we need any transformations? should we include polynomial terms? interactions?). In this problem, you will build an appropriate multiple regression model to analyze these data. **Bear in mind that your goal in the end will be to make inferential statements about the relationship between the explanatory variables and the response, including results of a hypothesis test for whether there is an association between smoking status and FEV after controlling for any other relevant covariates.**

i. Plot the data

At a minimum, make a pairs plot. I'd also suggest at least one other plot involving three or four variables in a single plot, either using facetting or colors (or both). Make whatever plots you think are informative. If it seems like transformations are necessary, you might consider exploring transformations in this exploratory phase before actually fitting models.

```r
fev_data <- fev_data %>%
  select(sex, smoke, age, height, fev)
ggpairs(fev_data)
```

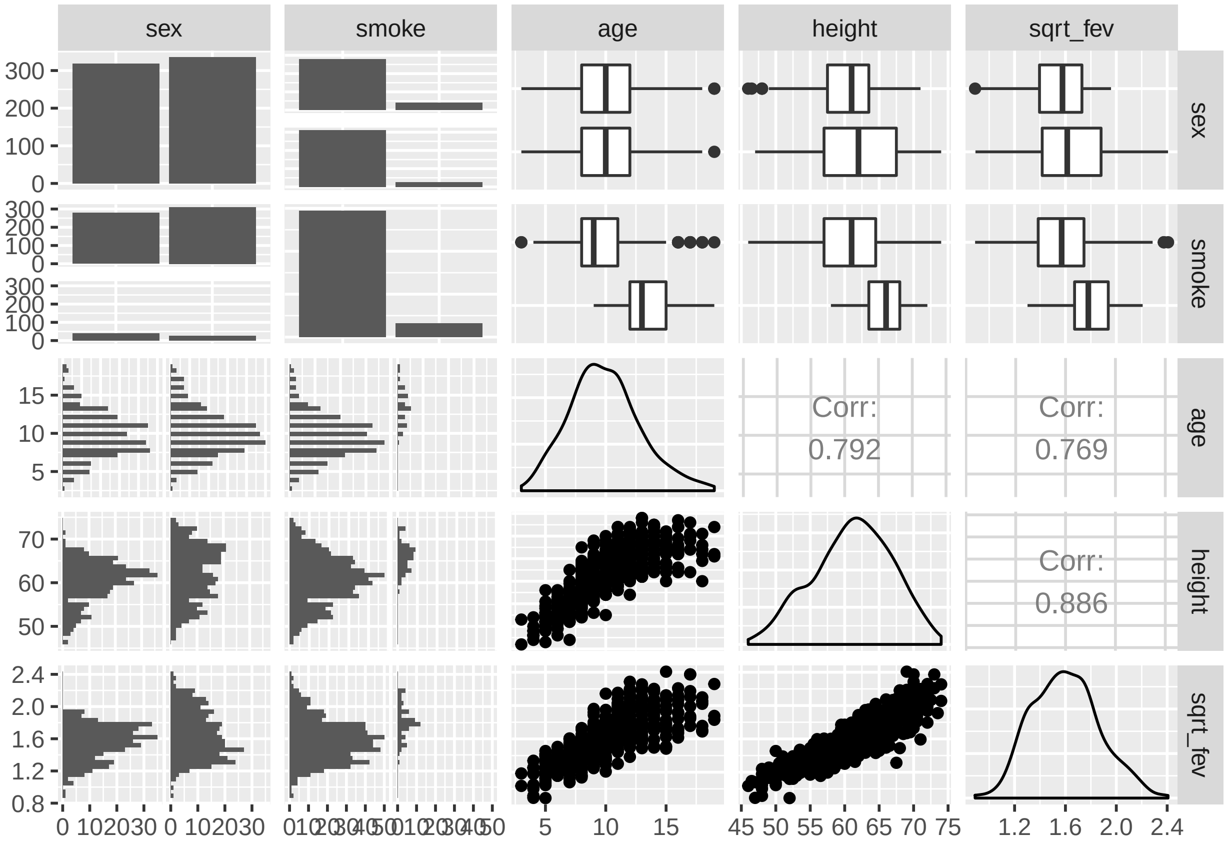
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The plots above suggest that a transformation of at least fev might be helpful. There is increasing variability of fev around the trend in both the plot against age and height.

```
fev_trans <- fev_data %>%
  transmute(
    sex = sex,
    smoke = smoke,
    age = age,
    height = height,
    sqrt_fev = sqrt(fev)
  )
ggpairs(fev_trans)
```

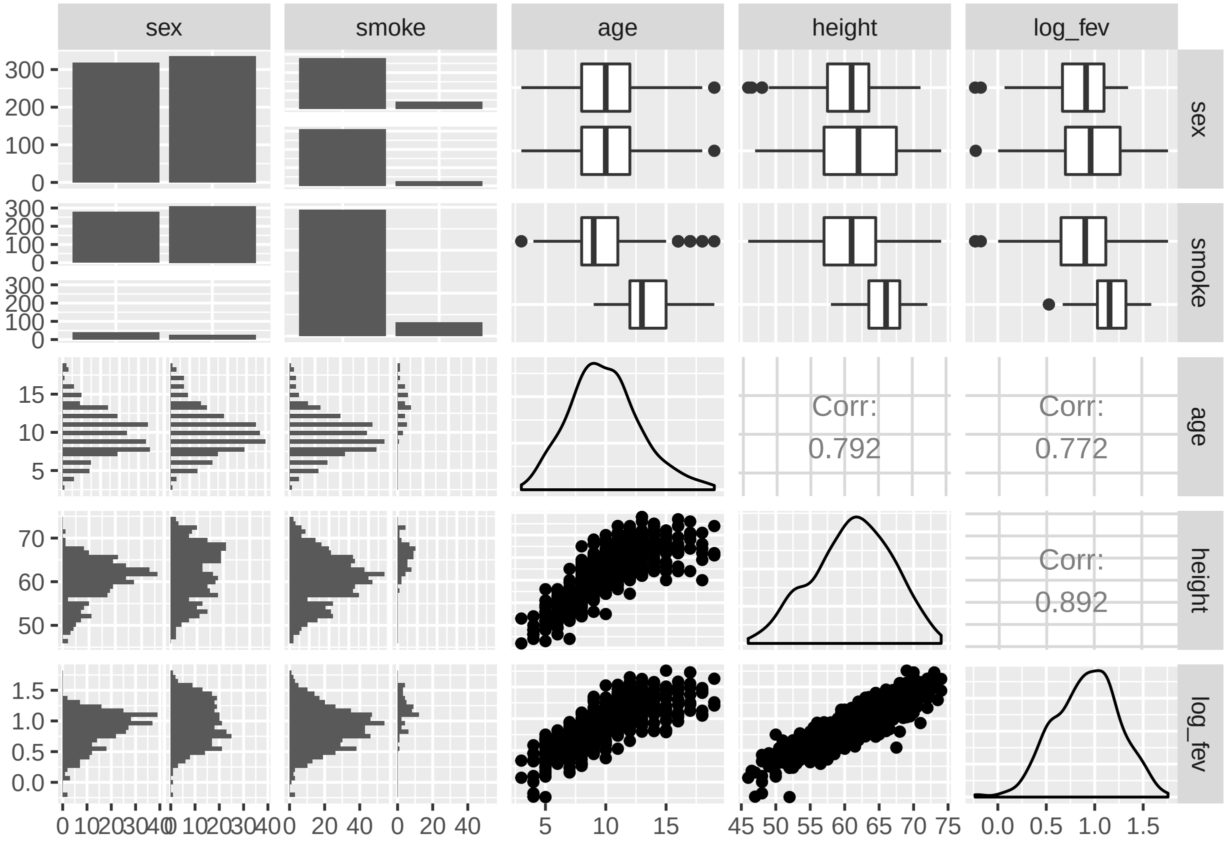
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Things look better with a square root transformation, but perhaps still not perfect. I still see more vertical spread for high values of age and height than for low values of age and height.

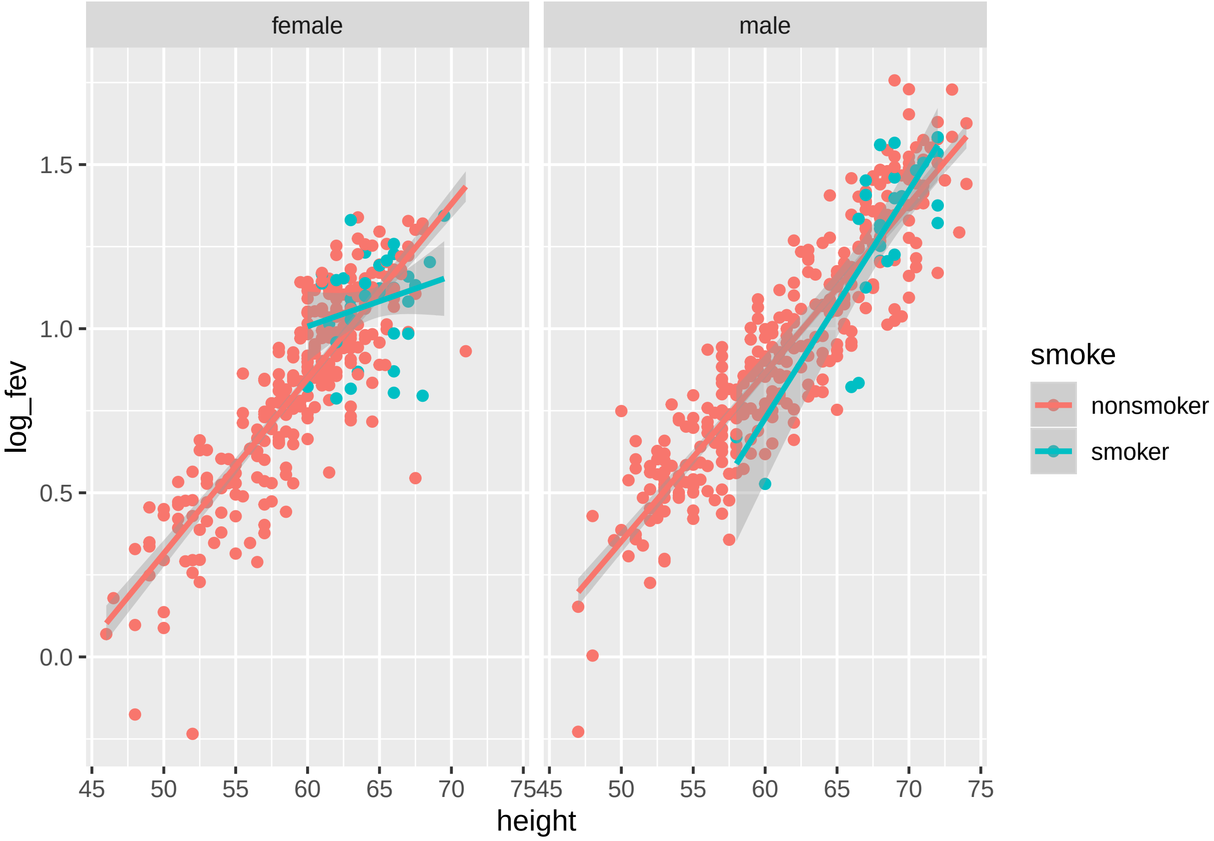
```
fev_trans <- fev_data %>%
  transmute(
    sex = sex,
    smoke = smoke,
    age = age,
    height = height,
    log_fev = log(fev)
  )
ggpairs(fev_trans)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

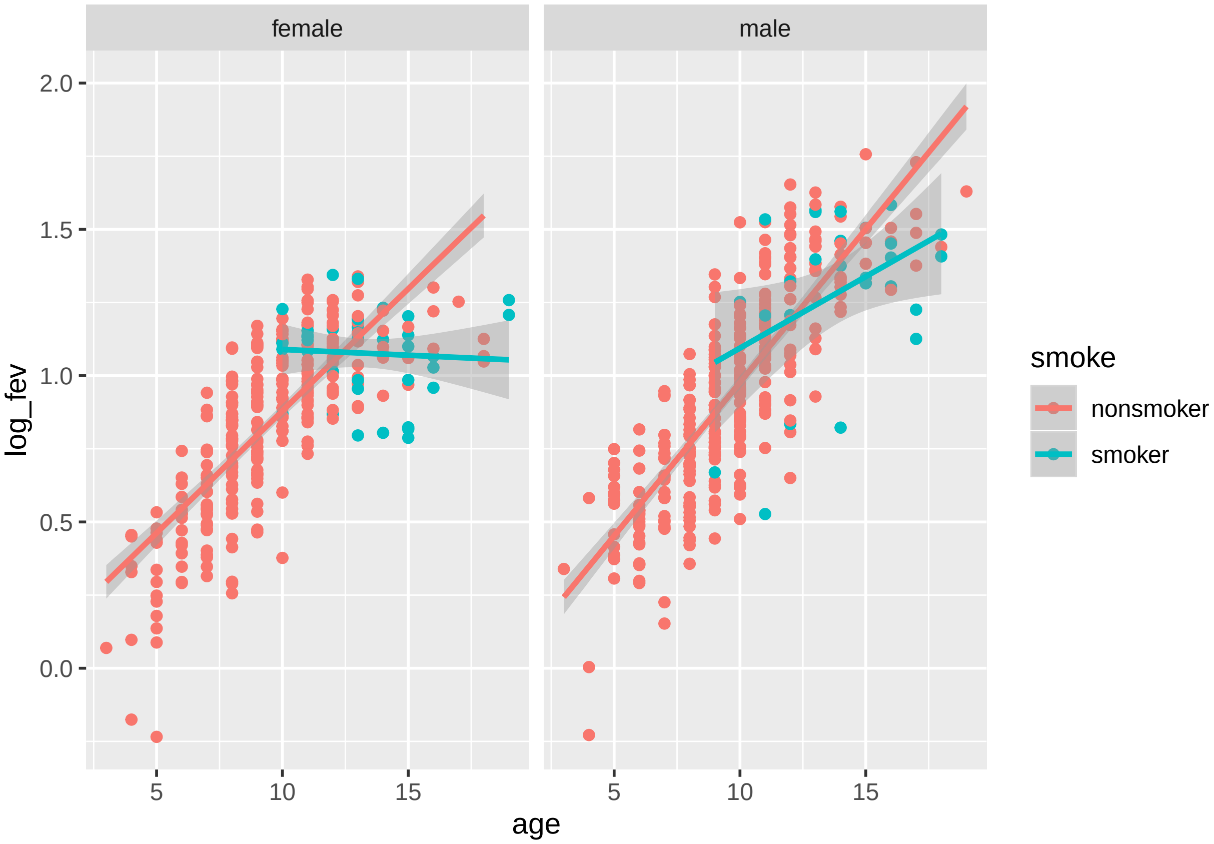


The plots above using a log transformation look better. There is a non-linear relationship between `age` and `log_fev`, but I could handle that using a polynomial term in `age`.

```
ggplot(data = fev_trans, mapping = aes(x = height, y = log_fev, color = smoke)) +
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap( ~ sex)
```



```
ggplot(data = fev_trans, mapping = aes(x = age, y = log_fev, color = smoke)) +
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap( ~ sex)
```



The plot faceting by sex suggests that an interaction between smoking status and age may be helpful since the lines for smokers and non-smokers have different slopes. It is possible that an interaction between sex and height could also be helpful.

ii. Find a model

Next, select an appropriate model for these data. Unless you have other ideas from your exploratory plots above, I suggest starting with a model that uses `fev` (or possibly a transformed version of `fev`) as the response and all other variables as explanatory variables, with no interactions or polynomial terms. Examine diagnostic plots and hypothesis tests, and consider the following possibilities for modifications to your initial model:

- transformations of the response and/or explanatory variables
- removing explanatory variables that don't seem helpful
- adding polynomial terms for variables that show a non-linear relationship with the response
- adding interactions between explanatory variables

For each candidate model you consider, please include diagnostic plots of residuals and results from cross-validation to evaluate performance of your model. If you consider any transformations of the response, be sure to calculate cross-validated MSE for predictions on the original scale of the data so that you get an assessment of model performance that is comparable to any models that you fit with the original data. However, note that if a transformation fixes problems with model residuals, you may prefer to use that transformation even if it results in lower scores for predictive performance. In your final document, I would like to see a coherent sequence of related models leading to your final model.

Unless you already zeroed in on a good model specification through your exploratory plots, you will likely need to try at least 3 models before you find a good choice, and possibly more. In this process, you may identify more than one model that seems plausible; on the other hand, you may find it difficult to obtain a model that seems exactly right. That's OK.

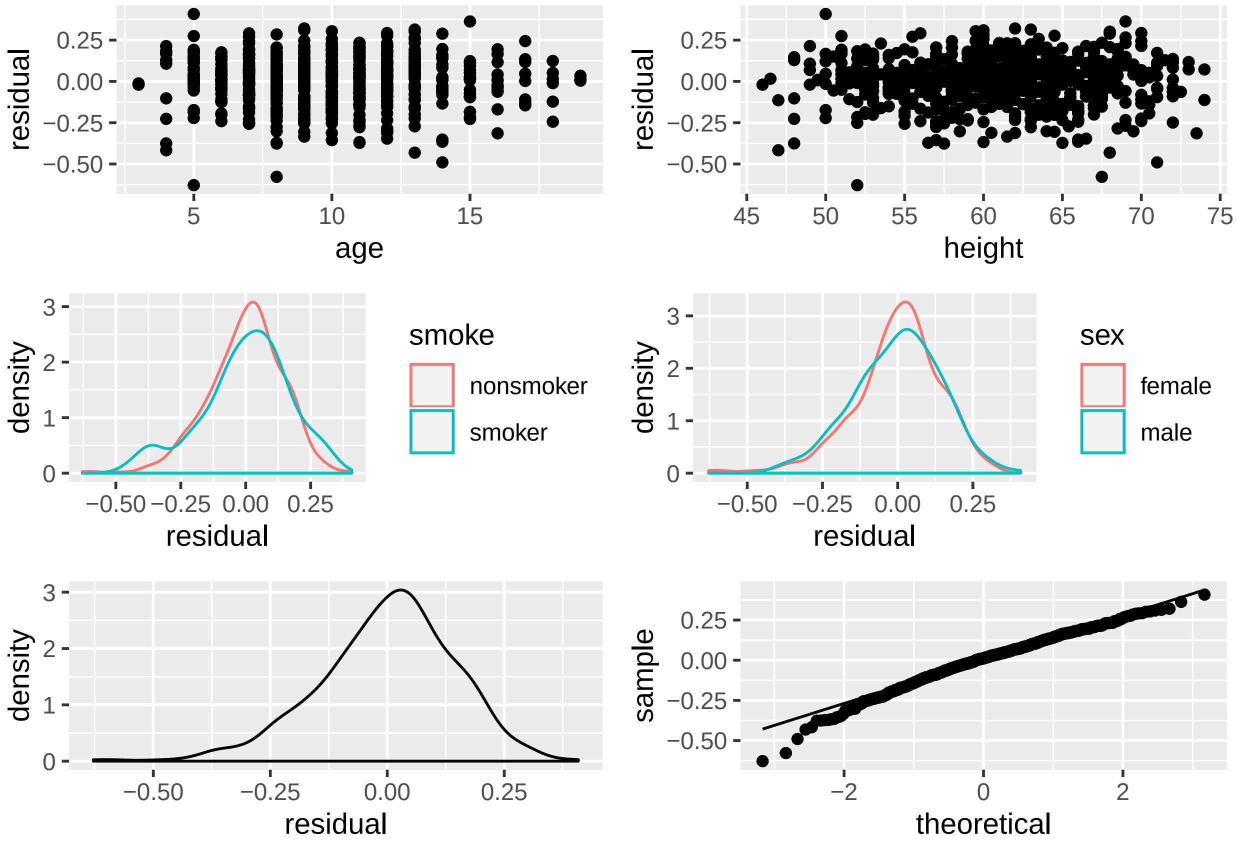
Model 1

I will start with a model that includes a quadratic term in age and an interaction between age and smoke, then use cross-validation to evaluate and see if I can drop any unnecessary terms.

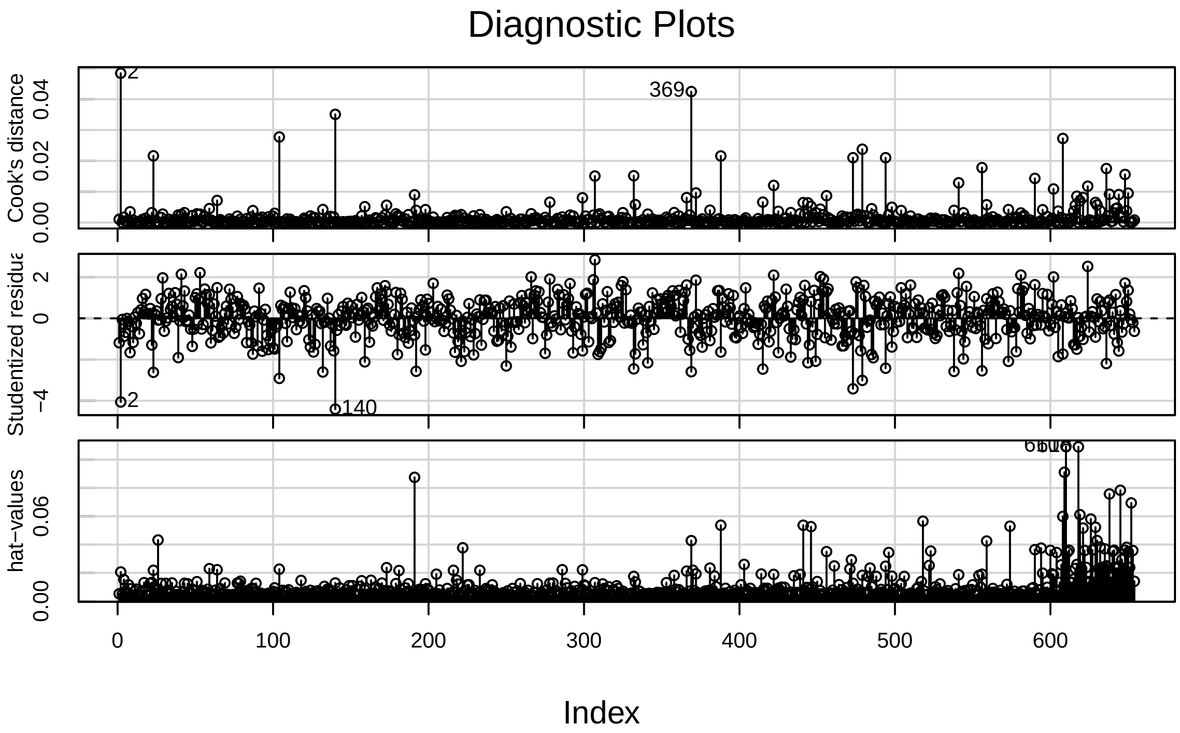
```r
fit1 <- lm(log_fev ~ age * smoke + sex + height + I(age^2), data = fev_trans)
summary(fit1)
```

```
##
## Call:
## lm(formula = log_fev ~ age * smoke + sex + height + I(age^2),
##     data = fev_trans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62823 -0.08784  0.01163  0.09641  0.40828
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.922e+00  8.381e-02 -22.929  < 2e-16 ***
## age              2.664e-02  1.325e-02   2.010  0.04485 *
## smokesmoker      1.039e-01  1.205e-01   0.863  0.38853
## sexmale          3.099e-02  1.183e-02   2.619  0.00903 **
## height           4.200e-02  1.898e-03  22.124  < 2e-16 ***
## I(age^2)        -5.687e-05  5.714e-04  -0.100  0.92076
## age:smokesmoker -1.134e-02  9.079e-03  -1.249  0.21204
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1455 on 647 degrees of freedom
## Multiple R-squared:  0.8112, Adjusted R-squared:  0.8094
## F-statistic: 463.3 on 6 and 647 DF,  p-value: < 2.2e-16
```

```r
fev_trans <- fev_trans %>% mutate(
  residual = residuals(fit1)
)

p1 <- ggplot(data = fev_trans, mapping = aes(x = age, y = residual)) +
  geom_point()

p2 <- ggplot(data = fev_trans, mapping = aes(x = height, y = residual)) +
  geom_point()

p3 <- ggplot(data = fev_trans, mapping = aes(x = residual, color = smoke)) +
  geom_density()

p4 <- ggplot(data = fev_trans, mapping = aes(x = residual, color = sex)) +
  geom_density()

p5 <- ggplot(data = fev_trans, mapping = aes(x = residual)) +
  geom_density()

p6 <- ggplot(data = fev_trans, mapping = aes(sample = residual)) +
  geom_qq() +
  geom_qq_line()

grid.arrange(p1, p2, p3, p4, p5, p6)
```



```r
car::influenceIndexPlot(fit1,
  vars = c("Cook", "Studentized", "hat"))
```



```r
2 * length(coef(fit1)) / nrow(fev_trans) # threshold for when we have to worry about leverage ("hat-val
```

```
## [1] 0.02140673
```

```r
set.seed(87362)
val_folds <- caret::createFolds(fev_data$fev, k = 10)
val_mse <- rep(NA, 10)

for(i in seq_len(10)) {
  fev_transformed_train <- fev_trans %>% slice(-val_folds[[i]])
  fev_transformed_val <- fev_trans %>% slice(val_folds[[i]])
  fev_val <- fev_data %>% slice(val_folds[[i]])

  lm_fit <- lm(log_fev ~ age * smoke + sex + height + I(age^2), data = fev_trans)

  y_hat_trans <- predict(lm_fit, newdata = fev_transformed_val)
  val_mse[i] <- mean((fev_val$fev - exp(y_hat_trans))^2)
}
mean(val_mse)
```

```
## [1] 0.1528427
```

Model 2

The plots and model summary output above made me think that:

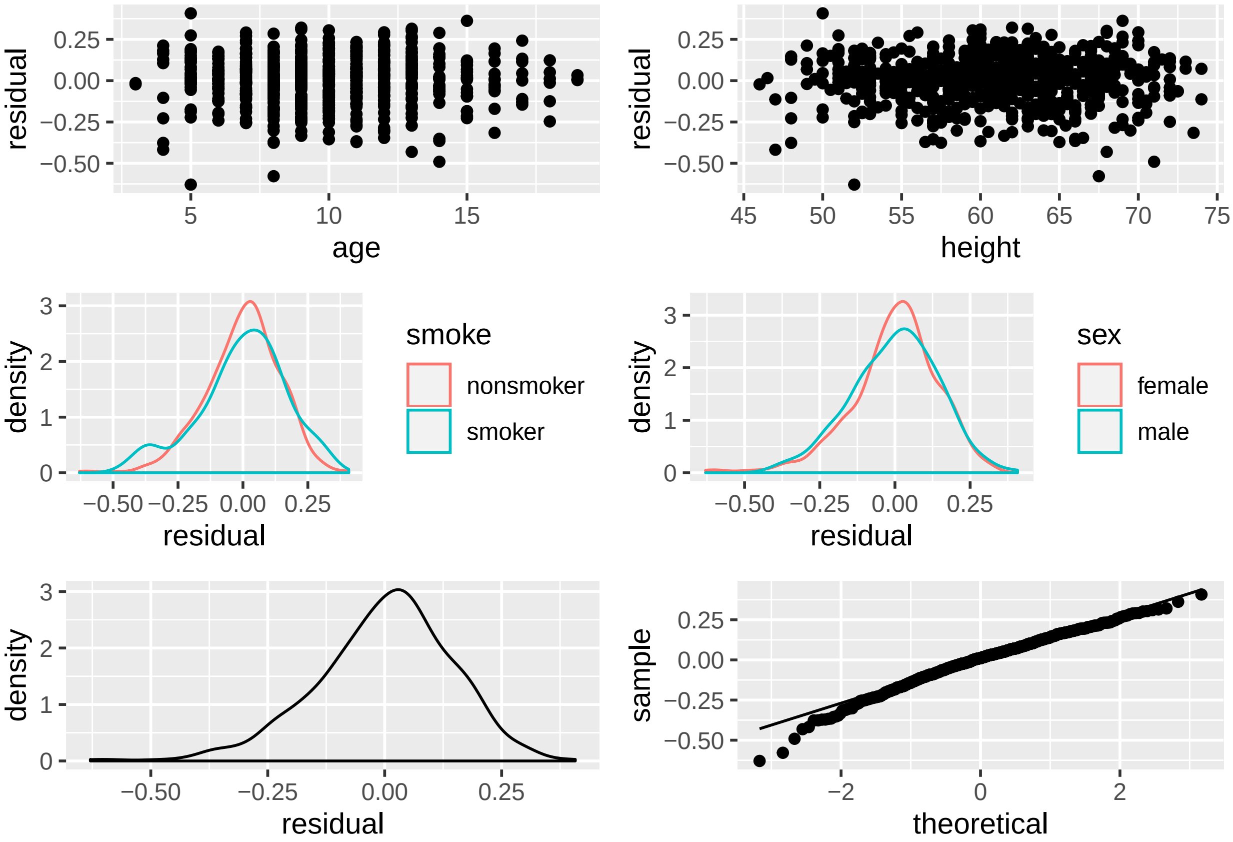
- I did not need the polynomial term in age.

```r
fit2 <- lm(log_fev ~ age * smoke + sex + height, data = fev_trans)
summary(fit2)
```

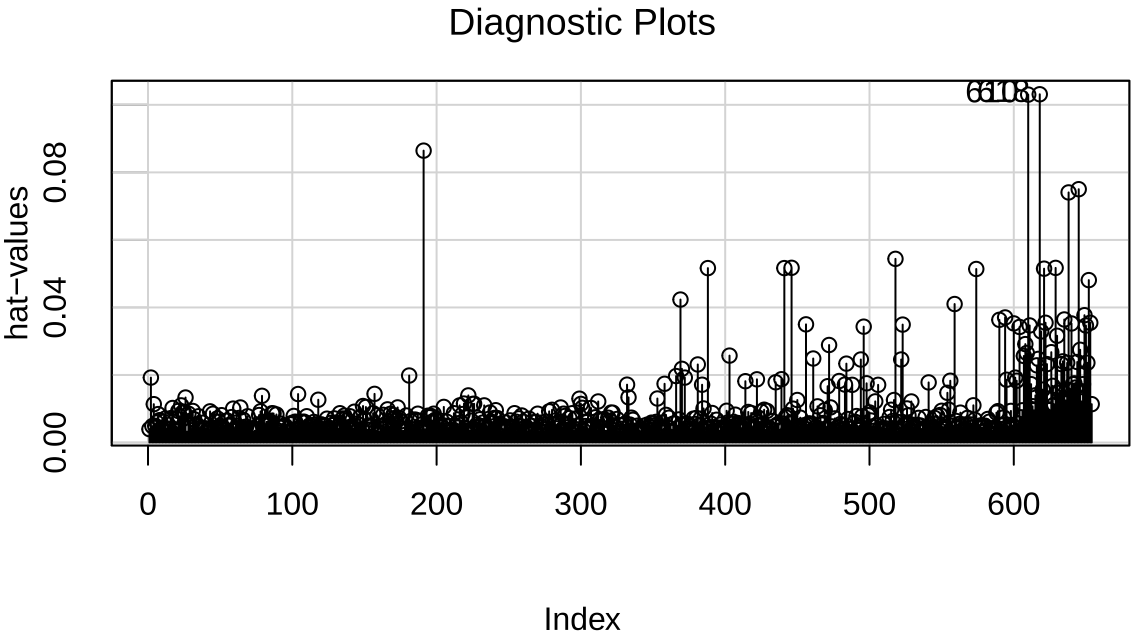
```
##
## Call:
```

```
## lm(formula = log_fev ~ age * smoke + sex + height, data = fev_trans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62926 -0.08783  0.01136  0.09658  0.40751
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.919494   0.080571 -23.824  < 2e-16 ***
## age              0.025368   0.003642   6.966 8.03e-12 ***
## smokesmoker      0.107884   0.113646   0.949  0.34282
## sexmale          0.030871   0.011764   2.624  0.00889 **
## height           0.042066   0.001759  23.911  < 2e-16 ***
## age:smokesmoker -0.011666   0.008465  -1.378  0.16863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1454 on 648 degrees of freedom
## Multiple R-squared:  0.8112, Adjusted R-squared:  0.8097
## F-statistic: 556.8 on 5 and 648 DF,  p-value: < 2.2e-16
```

```r
fev_trans <- fev_trans %>% mutate(
  residual = residuals(fit2)
)

p1 <- ggplot(data = fev_trans, mapping = aes(x = age, y = residual)) +
  geom_point()

p2 <- ggplot(data = fev_trans, mapping = aes(x = height, y = residual)) +
  geom_point()

p3 <- ggplot(data = fev_trans, mapping = aes(x = residual, color = smoke)) +
  geom_density()

p4 <- ggplot(data = fev_trans, mapping = aes(x = residual, color = sex)) +
  geom_density()

p5 <- ggplot(data = fev_trans, mapping = aes(x = residual)) +
  geom_density()

p6 <- ggplot(data = fev_trans, mapping = aes(sample = residual)) +
  geom_qq() +
  geom_qq_line()

grid.arrange(p1, p2, p3, p4, p5, p6)
```



```
car::influenceIndexPlot(fit2,
  vars = c("hat"))
```



```
2 * length(coef(fit2)) / nrow(fev_trans) # threshold for when we have to worry about leverage ("hat-val
```

```
## [1] 0.01834862
```

```
set.seed(87362)
val_folds <- caret::createFolds(fev_data$fev, k = 10)
val_mse <- rep(NA, 10)

for(i in seq_len(10)) {
  fev_transformed_train <- fev_trans %>% slice(-val_folds[[i]])
  fev_transformed_val <- fev_trans %>% slice(val_folds[[i]])
  fev_val <- fev_data %>% slice(val_folds[[i]])

  lm_fit <- lm(log_fev ~ age * smoke + sex + height, data = fev_trans)

  y_hat_trans <- predict(lm_fit, newdata = fev_transformed_val)
  val_mse[i] <- mean((fev_val$fev - exp(y_hat_trans))^2)
}
mean(val_mse)
```

```
## [1] 0.1528693
```

Removing the polynomial term resulted in a barely noticeable change in cross-validated MSE and no identifiable change in the estimated coefficients for the other terms in the model.

Model 3

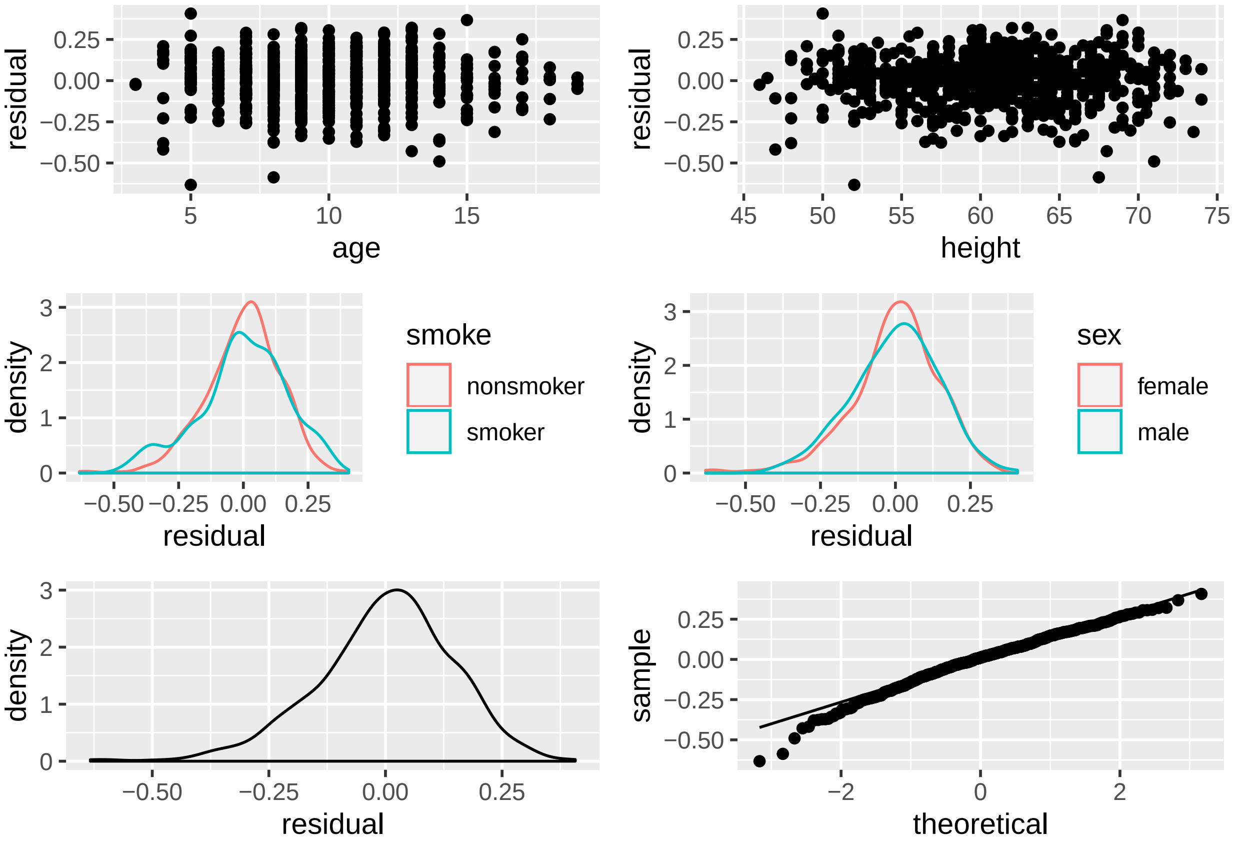
The interaction between age and smoking status is not statistically significant in the model output above either. I will try removing that term from the model.

```
fit3 <- lm(log_fev ~ smoke + sex + height + age, data = fev_trans)
summary(fit3)
```

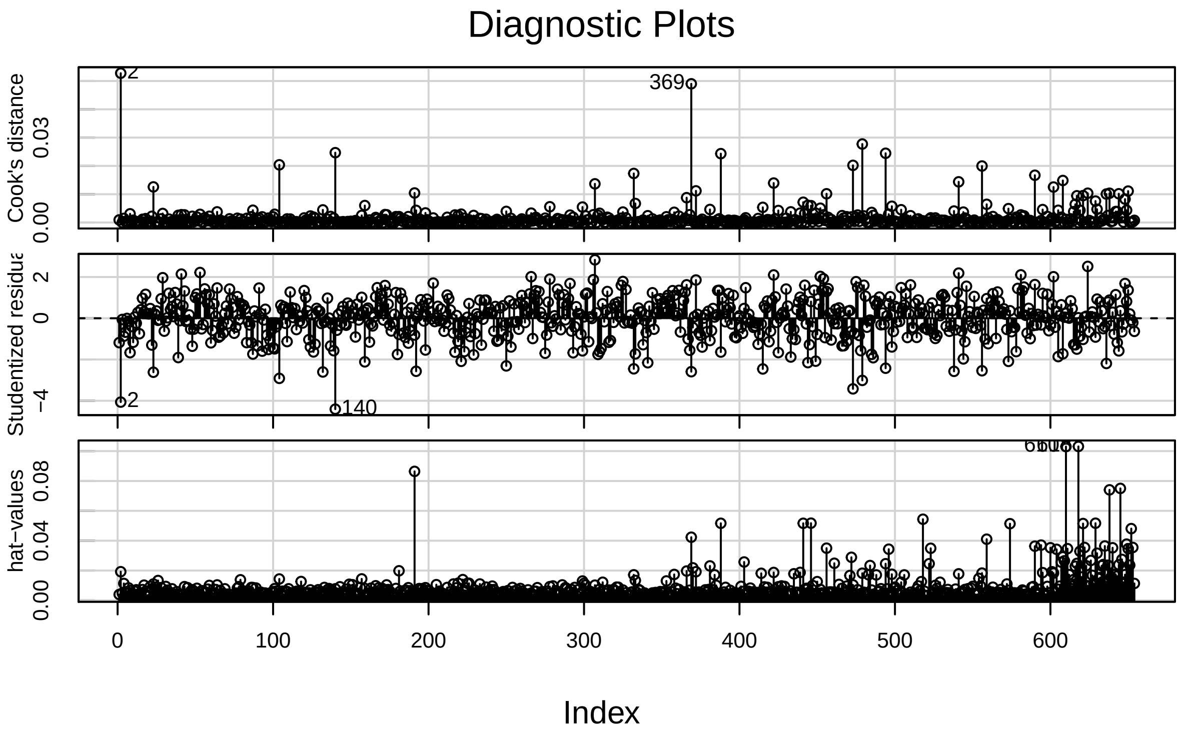
```
##
## Call:
## lm(formula = log_fev ~ smoke + sex + height + age, data = fev_trans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63278 -0.08657  0.01146  0.09540  0.40701
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.943998   0.078639 -24.721  < 2e-16 ***
## smokesmoker -0.046068   0.020910  -2.203   0.0279 *
## sexmale      0.029319   0.011719   2.502   0.0126 *
## height       0.042796   0.001679  25.489  < 2e-16 ***
## age          0.023387   0.003348   6.984  7.1e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1455 on 649 degrees of freedom
## Multiple R-squared:  0.8106, Adjusted R-squared:  0.8095
## F-statistic: 694.6 on 4 and 649 DF,  p-value: < 2.2e-16
```

```
fev_trans <- fev_trans %>% mutate(
  residual = residuals(fit3)
)
```

```r
p1 <- ggplot(data = fev_trans, mapping = aes(x = age, y = residual)) +
  geom_point()

p2 <- ggplot(data = fev_trans, mapping = aes(x = height, y = residual)) +
  geom_point()

p3 <- ggplot(data = fev_trans, mapping = aes(x = residual, color = smoke)) +
  geom_density()

p4 <- ggplot(data = fev_trans, mapping = aes(x = residual, color = sex)) +
  geom_density()

p5 <- ggplot(data = fev_trans, mapping = aes(x = residual)) +
  geom_density()

p6 <- ggplot(data = fev_trans, mapping = aes(sample = residual)) +
  geom_qq() +
  geom_qq_line()

grid.arrange(p1, p2, p3, p4, p5, p6)
```



```r
car::influenceIndexPlot(fit2,
  vars = c("Cook", "Studentized", "hat"))
```



```
2 * length(coef(fit3)) / nrow(fev_trans) # threshold for when we have to worry about leverage ("hat-val
```

```
## [1] 0.01529052
```

```
set.seed(87362)
val_folds <- caret::createFolds(fev_data$fev, k = 10)
val_mse <- rep(NA, 10)

for(i in seq_len(10)) {
  fev_transformed_train <- fev_trans %>% slice(-val_folds[[i]])
  fev_transformed_val <- fev_trans %>% slice(val_folds[[i]])
  fev_val <- fev_data %>% slice(val_folds[[i]])

  lm_fit <- lm(log_fev ~ smoke + sex + height + age, data = fev_trans)

  y_hat_trans <- predict(lm_fit, newdata = fev_transformed_val)
  val_mse[i] <- mean((fev_val$fev - exp(y_hat_trans))^2)
}
mean(val_mse)
```

```
## [1] 0.1535187
```

Removing the interaction resulted in a slightly larger increase in MSE, but this was still only a 0.4% increase in MSE.

Model 4

The plots of residuals vs explanatory variables and the QQ plot of residuals don't suggest any serious problems. All Cook's distances are OK too. There were a couple of observations with low studentized residuals, and several observations with high leverage. However, overall after using a log transformation the residuals are fairly close to normally distributed so I am not too concerned with the low studentized residuals.

Many observations are flagged as having high leverage. However, the two observations with highest leverage are children with young ages, and referring back to our original exploratory plots I am not concerned about any observations that don't fit the trend in that area. I didn't previously give you the code to do this, but we can also look at some of the other observations with high leverage using the code below. We see that these are mostly older females and younger male smokers. Again, referring to the exploratory plots above suggests that these people fit the trends in the rest of the data so I am not too concerned.

```
fev_data %>% slice(118, 222)
```

```
## # A tibble: 2 x 5
##   sex    smoke       age height   fev
##   <fct>  <fct>     <dbl>  <dbl> <dbl>
## 1 female nonsmoker     5   46.5  1.20
## 2 female nonsmoker     3   46    1.07
```

```
fev_data %>%
  mutate(
    leverage = hatvalues(fit3)
  ) %>%
  filter(leverage >= 0.025)
```

```
## # A tibble: 10 x 6
##    sex    smoke       age height   fev leverage
##    <fct>  <fct>     <dbl>  <dbl> <dbl>    <dbl>
##  1 male   smoker       12   72    3.75   0.0257
##  2 male   smoker       10   68    3.50   0.0263
##  3 male   smoker       11   72    4.64   0.0301
##  4 female nonsmoker    18   66    2.91   0.0255
##  5 female smoker       19   66    3.52   0.0312
##  6 female smoker       19   65.5  3.34   0.0323
##  7 female nonsmoker    18   64.5  3.08   0.0286
##  8 female nonsmoker    17   62    3.5    0.0274
##  9 male   smoker       18   67    4.09   0.0276
## 10 female nonsmoker    18   60    2.85   0.0414
```

Nevertheless, we can try fitting a model without the observations identified in the plots as having the most extreme residuals and leverages and see what changes. Here I'm removing the cases identified as potentially problematic in either model 2 or model 3. None of the results change substantially

```
fit4 <- lm(log_fev ~ age * smoke + sex + height,
  data = fev_trans %>% slice(-c(118, 222, 473, 140, 2, 610, 618)))
summary(fit4)
```

```
##
## Call:
## lm(formula = log_fev ~ age * smoke + sex + height, data = fev_trans %>%
##     slice(-c(118, 222, 473, 140, 2, 610, 618)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43634 -0.08835  0.01161  0.09125  0.41012
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.961393   0.079145 -24.782  < 2e-16 ***
## age              0.023139   0.003553   6.512  1.5e-10 ***
## smokesmoker      0.085488   0.118538   0.721   0.4711
## sexmale          0.023627   0.011474   2.059   0.0399 *
## height           0.043220   0.001724  25.064  < 2e-16 ***
## age:smokesmoker -0.010140   0.008913  -1.138   0.2557
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1408 on 641 degrees of freedom
## Multiple R-squared:  0.8175, Adjusted R-squared:  0.8161
## F-statistic: 574.3 on 5 and 641 DF,  p-value: < 2.2e-16
```

iii. Justify your model choice

In a few sentences, explain why you chose your selected model. Please discuss all diagnostic plots for your chosen model and what they say about the conditions for the regression model, as well as your results from cross-validation.

I found two reasonable models: model 2 above and model 3. Both models show consistent variability of residuals across different ages and heights, and a distribution of residuals that is fairly close to normal with a few low outliers that are not too severe. We investigated a few observations flagged as having large residuals or high leverage, and our model results did not change substantially depending on whether or not those observations were included.

iv. Explain your findings

Base on your selected model, discuss what your analysis has to say about the relationship between each of the four explanatory variables and FEV. This discussion should include interpretions of the coefficient estimates and confidence intervals for terms in your final model fit. Also discuss results for a hypothesis test about whether there is an association between smoking status and FEV after controlling for the other covariates. In writing this up, target a public health researcher who is not an expert in statistics.

```r
confint(fit2)
```

```
##                        2.5 %       97.5 %
## (Intercept)     -2.077704757 -1.761282822
## age              0.018216850  0.032518764
## smokesmoker     -0.115275604  0.331043729
## sexmale          0.007770161  0.053972332
## height           0.038611886  0.045521063
## age:smokesmoker -0.028287932  0.004956057
```

```r
confint(fit3)
```

```
##                    2.5 %       97.5 %
## (Intercept) -2.098414941 -1.789581413
## smokesmoker -0.087127344 -0.005007728
## sexmale      0.006308481  0.052330236
## height       0.039498923  0.046092655
## age          0.016812109  0.029962319
```

Depending on the model selected, we found either no evidence or only weak evidence of an association between smoking status and log of FEV after accounting for age, sex, and height. In the model that allowed for the relationship between age and FEV to depend on smoking status by including an interaction, the p-value for a test of whether smoking status had an effect on FEV was 0.34, indicating no evidence of an association. In the model that had the same relationship between age and FEV for both smokers and non-smokers, the p-value was 0.027, indicating only some weak evidence of an association between smoking and FEV. Based on this second model, we are 95% confident that on average, a child who smokes has log FEV values between 0.087 and 0.005 units lower than a child of the same age, height, and sex who does not smoke. However, in the model with an interaction between age and smoking status, our 95% confidence interval included values suggesting that this difference could plausibly be either positive or negative.

All other estimates were consistent across the two models. We are 95% confident that on average, a male has a log FEV between about 0.006 and 0.052 units higher than a female of the same age, height and smoking status; that a 1 cm increase in height while holding fixed sex, age, and smoking status is associated with an increase in log FEV of between about 0.039 and 0.046 units; and that a 1 year increase in age while holding fixed sex, height, and smoking status is associated with an increase in log FEV of between about 0.017 and 0.030 units.

v. Reflect

Do your conclusions from the univariate analysis in part (a) agree with your conclusions from the multiple regression analysis here? If your conclusions are the same in both parts, explain why (why didn't your results change when you added more variables?). If your conclusions are different, explain why they are different (why did your results change when you added more variables?). Consider the variables being analyzed, the context of the data, and your exploratory plots in part (b) i.

The conclusions from parts (a) and (b) are different. In part (a) we found strong evidence that smokers had higher FEV than non-smokers, but in part (b) we found either no evidence of an association between smoking status and FEV after accounting for smoking status, or weak evidence that smokers had lower FEV on average than non-smokers.

Based on our plots, the reason for this difference in conclusions appears to be that in general smokers were older and taller than non-smokers. When we analyzed the data without controlling for age and height, the effects of age and height on FEV got mixed up with the effect of smoking status on FEV, making it appear that smoking was associated with increased FEV when in fact this was just due to differences in age and heigh between smokers and non-smokers in our sample. By including age and height as variables in our model, we are able to isolate the effect of smoking on FEV for children of a similar age and height.

Conceptual Problems

If you prefer, you can write your answers to all conceptual problems by hand and turn in a physical copy. It's also fine if you want to write your answers up in LaTeX and push the pdf to GitHub.

Problem 2: Pre-planning analyses is difficult.

It is sometimes claimed that we should not look at our data before formulating our models and a specific plan for any inference tasks we will carry out (such as any hypothesis tests we want to conduct). Let's consider this in the context of Problem 1.

(a) Suppose we pre-planned to conduct an ANOVA analysis with FEV as the response and smoking status as the explanatory variable, as in part 1 (a) on problem set 4. Would we have drawn the correct conclusion about the relationship between smoking and lung health from that analysis? A 1 or 2 sentence response is fine.

We would not have drawn the correct conclusion about the relationship bretween smoking and lung health from a simple ANOVA analysis with only smoking status as the explanatory variable. That analysis found a "statistically significant" association between smoking status and lung health, but in the wrong direction. The analysis found that smoking led to healther lungs, when in fact it is associated with less healthy lungs after controlling for other important variables such as the subject's height.

(b) Suppose we pre-planned our analysis, but we knew we should account for the effects of other related variables such as age, height, and sex. Do you think it's plausible that we could have guessed at a reasonably correct model specification without an exploratory process that involved looking at plots of the data? In light of the fact that hypothesis tests and confidence intervals from misspecified models do not have any guarantees of correct performance (in terms of correct probabilities of Type I Error or correct coverage rate for confidence intervals), would you be confident in conclusions from such an analysis? A 1 or 2 sentence response is fine.

It would have been difficult to correctly guess a reasonable model specification without looking at plots of the data. I definitely would not have been able to guess that we would need a quadratic term in height or to take a log transformation of the response variable without examining plots. Perhaps an expert in lung health would have been able to guess that, but that would only have come from previous experience with looking at plots of similar data. This means that any inferences from a model that was specified without looking at the data would likely be based on an incorrectly specified model, and therefore unreliable.

Problem 3: You can always find results if you look hard enough, part 1

The point of this problem is that you can misuse statistics to "prove" anything you want as long as you have enough possibilities for your analysis and you look at the data hard enough. Go to https://projects.fivethirtyeight.com/p-hacking/

(a) Suppose you are a Democratic data analyst with an agenda. You want to show that there is a positive association between the amount of power held by Democrats in political office and the strength of the economy. Find a combination of settings for how you will measure the power of Democrats in office and how you will measure economic performance that "proves" you are right, at a statistically significant level. Note the settings you used and the p-value you achieved here; no need to write anything else.

I included Senators and Representatives in the politicians, GDP in the economic measures and excluded recessions. This gave a positively sloped line and a p-value of 0.01.

(b) Suppose you are a Republican data analyst with an agenda. You want to show that there is a positive association between the amount of power held by Republicans in political office and the strength of the economy. Find a combination of settings for how you will measure the power of Republicans in office and how you will measure economic performance that "proves" you are right, at a statistically significant level. Note the settings you used and the p-value you achieved here; no need to write anything else.

I included Governors and Senators in the politicians, Employment and Inflation in the economic measures, and excluded recessions. This gave a positively sloped line and a p-value of 0.01.

Problem 4: You can always find results if you look hard enough, part 2

The point of this problem is that if you conduct enough hypothesis tests, you will find a "statistically significant" result in your analysis.

(a) Recall that a Type I Error in a hypothesis test is when you reject the null hypothesis, but the null hypothesis is actually correct. Suppose you will reject the null hypothesis if you find a p-value that is less than the significance level $\alpha = 0.05$. What is the probability of making a Type I Error in this case, if the null hypothesis is correct? You can answer in 1 sentence.

If the null hypothesis is correct and we will declare a statistically significant result if the p-value is less than $\alpha = 0.05$, the probability of making a Type I Error is 0.05. That is, for 5% of samples, we will incorrectly decide the that null hypothesis is wrong.

(b) Check out this cartoon (I acknowledge that it's not funny, even though it's a cartoon) https://xkcd.com/882/. In the cartoon, the researchers conducted 20 different hypothesis tests and found a "statistically significant" result in one of them. How does this relate to your answer to part (a)? Do the scientists' results give us strong evidence that green beans cause acne? You can answer in 2 or 3 sentences.

If the null hypothesis were correct in all 20 hypothesis tests the researchers conducted, we would expect them to make a Type 1 error in 5% of those tests, or 1 of those tests. Finding one "statistically significant" result in 20 hypothesis tests is exactly what we would expect if the null hypothesis was true. This does not provide strong evidence that green jelly beans cause acne.

Problem 5. The garden of forking paths

Read this article about the "garden of forking paths": http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

In the section titled "Theoretical Framework", the authors propose a four-level hierarchy of analyses:

1. *'Simple classical test based on a unique test statistic, T, which when applied to the observed data yields $T(y)$.'* Less formally stated, the idea here is that there is only one conceivable hypothesis test that could be conducted to answer the scientific question interest. We conduct that test and report the result.
2. *'Classical test pre-chosen from a set of possible tests: thus, $T(y; \phi)$, with preregistered ϕ. For example, ϕ might correspond to choices of control variables in a regression, transformations, and data coding and excluding rules, as well as the decision of which main effect or interaction to focus on.'* The idea here is that there are many possible models we could fit (for example, with different transformations used and covariates included) and many different tests we could conduct, but we think carefully about our analysis and decide on what model we will fit and what hypothesis test we will conduct **before collecting data**.
3. *'Researcher degrees of freedom without fishing: computing a single test based on the data, but in an environment where a different test would have been performed given different data; thus $T(y; \phi(y))$, where the function $\phi(\mathring{u})$ is observed in the observed case.'* In this case, there are many possible models we could fit, but this time we choose our model specification and possibly the details of the test we conduct in a way that depends on the data we have observed. For example, we choose the transformations we use, inclusion of any polynomial terms, and interactions for the model based on an exploratory analysis of the data. We then conduct a single hypothesis test to answer the scientific question of interest based on the selected model.
4. *' "Fishing": computing $T(y; \phi_j)$ for $j = 1, \dots, J$: that is, performing J tests and then reporting the best result given the data, thus $T(y; \phi^{best(y)})$ '.* Once again there are many possible models and hypothesis tests we might conduct. However, this time we conduct many of these tests (J of them) and report only the results of the tests that support our claims or seem to be publishable.

(a) In the article's four-level hierarchy of analyses, where does our analysis of the data set in Problem 1 (a) fall? Explain in a sentence or two.

Our analysis of the data set in Problem 1 (a) was at level 1 of the hierarchy. In that analysis, we chose how we would perform our analysis (a t test from a one-way ANOVA model) before looking at the data.

(b) In the article's four-level hierarchy of analyses, where does our analysis of the data set in Problem 1 (b) fall? Explain in a sentence or two.

Our analysis of the data set in Problem 1 (b) was at level 3 of the hierarchy. In that analysis, we chose how we would perform our analysis (what data transformations and nonlinear functions of explanatory variables we would use) in a data-dependent way, by looking at our plots. Since our analysis depended on the data we had, it was not a pre-registered analysis, so we were not at level 2 of the hierarchy. However, we were not specifically fishing for a statistically significant result to report. Our goal was to obtain a model that described the data well, and to draw whatever inferences were appropriate based on that model; this means we were not at level 4 of the hierarchy.

(c) In the article's four-level hierarchy of analyses, where do the analyses in Problems 3 and 4 fall? Explain in a sentence or two.

The analyses in problems 3 and 4 of this problem set was at level 4 of the hierarchy. We tried many analyses based on different definitions of the explanatory and response variables, and reported only the results from one statistically significant model or test.

(d) Write a summary of the article that's about 2 paragraphs long. You should outline (a) some of the types of decisions that data analysis can make as part of their analysis and why that can affect our ability to trust the validity of published p-values; and (b) some of the authors' suggestions for how we (as the field of statistics) can proceed.

There are many different decisions that a researcher can make as part of their analysis. These include how the variables used in the model are defined, what data to include or exclude from an analysis, and details of the model specification such as data transformations, what quadratic or polynomial terms to include, and what interactions to include in a model. On one hand, it is often necessary to tune our analysis to the specific data set we are working with, since this is necessary to correctly specify a model and also lets us find out about relationships between the variables we are studying that we might not have thought of otherwise. On the other hand, this invalidates the hypothesis tests we might have conducted. For example, suppose that we are considering a model with a lot of possible explantory variables, and we want to allow for the possibility that there may be an interaction between some of those variables. If we try out all possible interactions between every pair of those variables, it is very likely that we will find an interaction that is "statistically significant"" (this is related to the multiple testing we explored in problem 3). This flexibility in our analysis means that the final model I settle on may include terms that are "statisticaly significant" as determined by their p-values, but those p-values came out of a multiple testing scenario and therefore cannot be taken at face value.

The authors' main suggestions for the field revolve around preregistering our analyses whenever possible. If we publicly state what our analysis plan will be, including all variables that will be included in the analysis, all transformations, the model statement, and any hypothesis tests that will be conducted before collecting and exploring the data, the p-values resulting from that analysis can be trusted more. In cases where the analysis cannot be planned in detail before the data are collected, it may sometimes be possible to conduct two different experiments in sequence. The first experiment will be used to conduct an exploratory, data-dependent, analysis; the p-values from this study will not be reliable, but this study will inform future work. Then the analysis procedure for a second experiment will be pre-registered, using what we learned from the first experiment; the p-values from this second analysis will be more reliable. In some fields, it is not possible to conduct new experiments to gather new data. In those fields, the authors propose that we conduct a broader set of analyses, from multiple models, and discuss the results from all of these models. This will give a more inclusive picture of the range of results we could obtain from reasonable analyses of the data.

Problem 6. MSE and Bias/Variance

The data set read in below has 221 observations from a light detection and ranging (LIDAR) experiment, with measurements of two variables:

- `range`: distance travelled before the light is reflected back to its source.
- `logratio`: logarithm of the ratio of received light from two laser sources.

References:

Sigrist, M. (Ed.) (1994). Air Monitoring by Spectroscopic Techniques (Chemical Analysis Series, vol. 197). New York: Wiley.

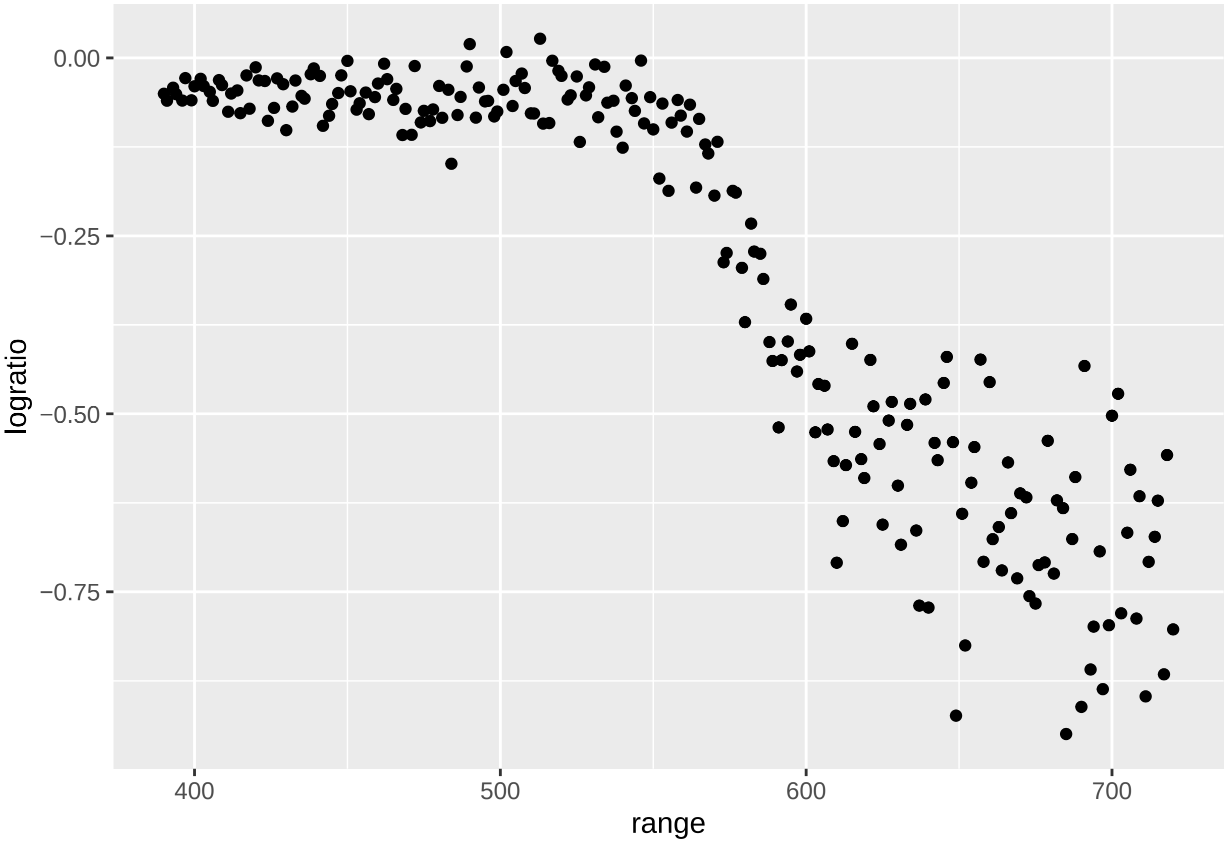
Ruppert, D., Wand, M.P. and Carroll, R.J. (2003) Semiparametric Regression Cambridge University Press. http://stat.tamu.edu/~carroll/semiregbook/

The description above was adapted from the documentation provided by Wasserman, L. (2007) All of Nonparametric Statistics. New York: Springer.

```r
lidar <- read_table2("http://www.evanlray.com/data/all-of-nonparametric-stat/lidar.dat")
```

```
## Parsed with column specification:
## cols(
##   range = col_double(),
##   logratio = col_double()
## )
```

```r
ggplot(data = lidar, mapping = aes(x = range, y = logratio)) +
  geom_point()
```



Please answer the following questions based on an understanding of the bias/variance trade-off and the relative strengths of these methods.

(a) For LIDAR data like these, would you expect a simple linear regression model or a degree 5 polynomial model to have a lower training set residual sum of squares? Explain your answer in a sentence or two.

The training set RSS will be lower for the degree 5 polynomial than for the degree 1 polynomial. The higher degree polynomial has more flexibility, so can find a fitted curve that is closer to the training data set. That means its residuals will be smaller, so the RSS will be smaller.

(b) For LIDAR data like these, would you expect a simple linear regression model or a degree 5 polynomial model to have a lower bias for predicted values at the test point $x_0 = 550$? Explain your answer in a sentence or two.

In general, more flexible models have lower bias than less flexible models. The degree 5 polynomial has lower bias than the degree 1 polynomial. In particular, at the test point $x_0 = 550$ I expect a line from a simple linear regression model to go well below the data, suggesting substantial bias for the linear regression model at that point. I would not expect the degree 5 polynomial model to have much bias at that test point.

(c) For LIDAR data like these, would you expect a simple linear regression model or a degree 5 polynomial model to have a lower variance for predicted values at the test point $x_0 = 550$? Explain your answer in a sentence or two.

In general the more flexible model will have higher variance than the less flexible model. The degree 5 polynomial model will have higher variance than the simple linear regression model.

(d) For LIDAR data like these, would you expect a simple linear regression or a degree 5 polynomial model to have a lower expected test set MSE at the test point $x_0 = 550$? Explain your answer in a sentence or two. (Note: Formally, it is not possible to be sure of the answer to this question without knowing the true data generating process, but you should have some intuition. I think there is a correct answer, but you will receive full credit as long as your answer is reasonable and you justify it with a correct discussion of the bias/variance tradeoff made by these models and how that relates to the specifics of this data set. Evan will grade this question.)

To answer this question, we need to think about the bias/variance trade-off made by each of these models in the context of this particular data set. We decided above that the simple linear regression model has higher bias than the degree 5 polynomial, but lower variance. The question is, which if these differences is larger?

My intuition is that at the particular point $x_0 = 550$, the bias of the simple linear regression model is so large that the squared bias will overwhelm any differences in variance. This is especially the case because our sample size is fairly large in this example, so that I think there would not be too much variation across different training sets in the fitted values from the degree 5 polynomial at $x_0 = 550$. Thus, I think that the degree 5 polynomial would have lower expected test set MSE at $x_0 = 550$.

Note that my answer would change at a different value of x_0, like $x_0 = 650$. Based on the plot above, I'd expect a fit from a simple linear regression to pass about through the middle of the observed data at that location. Since I believe that the true function we're estimating would pass through the data at that point, I think the bias of the simple linear regression model at $x_0 = 650$ would be minimal. In that case, the higher variance of the degree 5 polynomial might mean that overall, the degree 5 polynomial would have higher expected test set MSE at that point.

Collaboration and Sources

If you worked with any other students on this assignment, please list their names here.

If you referred to any sources (including our text book), please list them here. No need to get into formal citation formats, just list the name of the book(s) you used or provide a link to any online resources you used.