Homework #7

Dustin Leatherman

11/9/2019

8.11

a

Fit regression model (8.22)

```
brand <- read.csv("~/Downloads/brandperference.csv")
brand.m.full <- lm(liking ~ moisture * sweetness, data = brand)

brand.m.full %>% tidy %>% kable
```

term	estimate	std.error	statistic	p.value
(Intercept)	27.150	6.4648086	4.199660	0.0012326
moisture	5.925	0.8797490	6.734875	0.0000209
sweetness	7.875	2.0443520	3.852076	0.0023015
moisture:sweetness	-0.500	0.2782011	-1.797261	0.0974862

b

Test whether or not the interaction terms can be dropped from the model. use $\alpha = 0.05$. State the alternatives, decision rule, and conclusion.

$$H_0 : X_3 = 0 H_A : X_3 \neq 0$$

There is weak evidence that the interaction between moisture and sweetness have a significant impact on mean liking (two-tail t-test. p-value = 0.09749).

8.24

A tax consultant studied the current relation between selling price and assessed valuation of one-family residential dwellings in a large tax district by obtaining data for a random sample of 16 recent "arm's-length" sales transactions of one-family dewllings located on corner lots and for a random sample of 48 recent sales of one-family dwellings *not* located on corner lots. Assume that the error variances in the two populations are equal and that regression model (8.49) is appropriate.

a

Plot the sample data for the two populations as a symbolic scatter plot. Does the regression relation appear to be the same for the two populations?

```r
valuations <- read.csv("~/Downloads/assessedvaluations.csv")

valuations %>%
  ggpairs(
    columns = c("Price", "Value"),
    aes(shape = factor(CornerLot), color = factor(CornerLot), alpha = 0.5)
  )
```

The regression relation appears to be similar for Price vs Value. There is a positive correlation between Price and Value for Corner Lots and non-Corner lots though there is stronger correlation for non-corner lots.

b

> Test for identity of the regression functions for dwellings on corner lots and dwellings in other locations; control the risk of Type I error at 0.05. State the alternatives, decision rule, and conclusion.

```r
valuations.m.full <- lm(Price ~ Value * CornerLot, data = valuations)
valuations.m.no_x <- lm(Price ~ Value + CornerLot, data = valuations)

valuations.m.full %>% tidy %>% kable
```

term	estimate	std.error	statistic	p.value
(Intercept)	-126.905171	14.7224698	-8.619829	0.0000000
Value	2.775898	0.1962820	14.142397	0.0000000
CornerLot	76.021532	30.1313556	2.523004	0.0143040
Value:CornerLot	-1.107482	0.4055382	-2.730895	0.0082809

$H_0 : \beta_4 = 0$

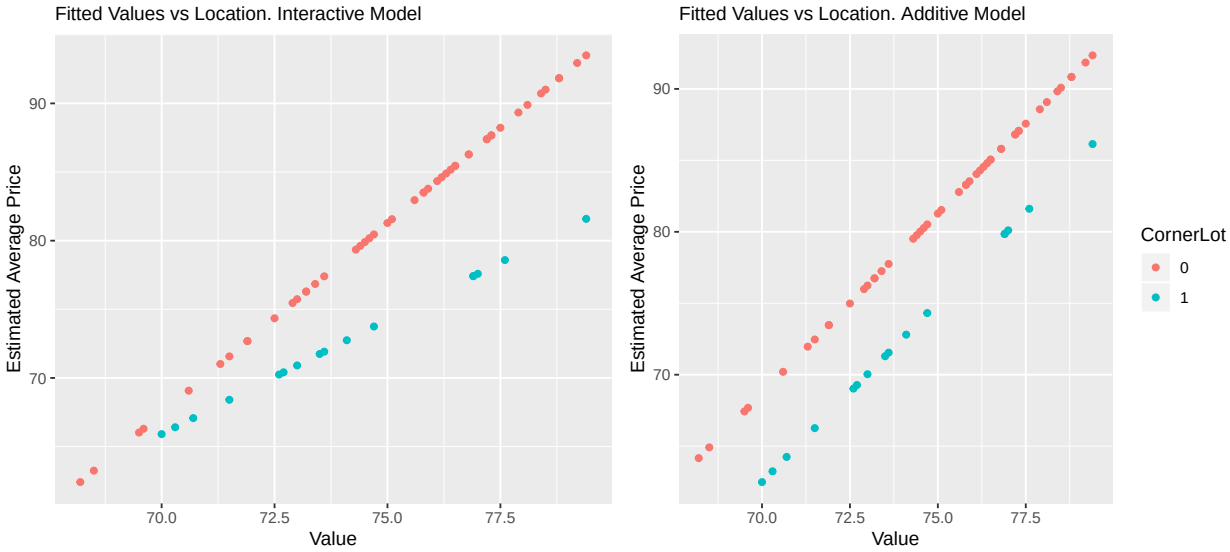
$H_A : \beta_4 \neq 0$

There is convincing evidence that the interaction between location and value has a significant effect on Price of a single-family dwelling (two-tailed t-test. p-value = 0.00828).

c

Plot the estimated regression functions for the two populations and describe the nature of the differences between them.

```
valuations.p.x <-
  valuations.m.full %>%
  ggplot(aes(x = Value, y = .fitted, color = factor(CornerLot))) +
  geom_point(show.legend = FALSE) +
  labs(color = "CornerLot", subtitle = "Fitted Values vs Location. Interactive Model") +
  ylab("Estimated Average Price")

valuations.p.additive <-
  valuations.m.no_x %>%
  ggplot(aes(x = Value, y = .fitted, color = factor(CornerLot))) +
  geom_point() +
  labs(color = "CornerLot", subtitle = "Fitted Values vs Location. Additive Model") +
  ylab("Estimated Average Price")

grid.arrange(valuations.p.x, valuations.p.additive, ncol = 2)
```



As expected, the key difference between the two models is the slope of the Value vs Fitted Price. The model containing the interaction term has differening slopes between CornerLots and non-Corner Lots while the Additive model has the same slopes for both.

9.15

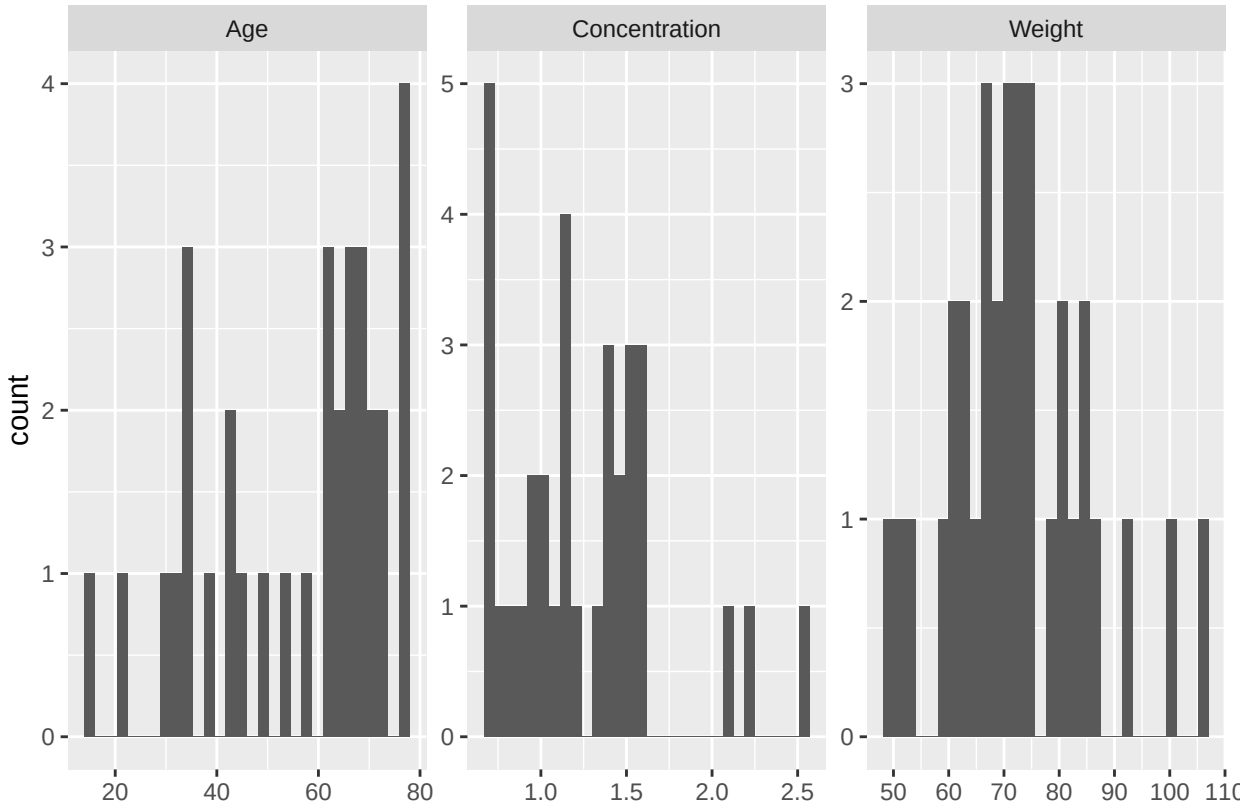
Creatinine clearance (Y) is an important measure of kidney function, but it is difficult to obtain in a clinical office setting because it requires 24-hour urine collection. To determine whether this measure can be predicted from some data that are easily available, a kidney specialist obtained the data that follow for 33 male subjects. The predictor variables are serum creatinine concentration, age, and weight.

a

Prepare separate histograms for each of the three predictor variables. Are there any noteworthy features in these plots? Comment.

```
kidney <- read.csv("~/Downloads/kidneyfunction.csv")
kidney %>%
  gather(c(-Clearance), key = "variable", value = "value") %>%
    ggplot(aes(x = value)) +
    geom_histogram() +
    facet_wrap(~ variable, ncol = 3, scales = "free") +
    xlab("")
```

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



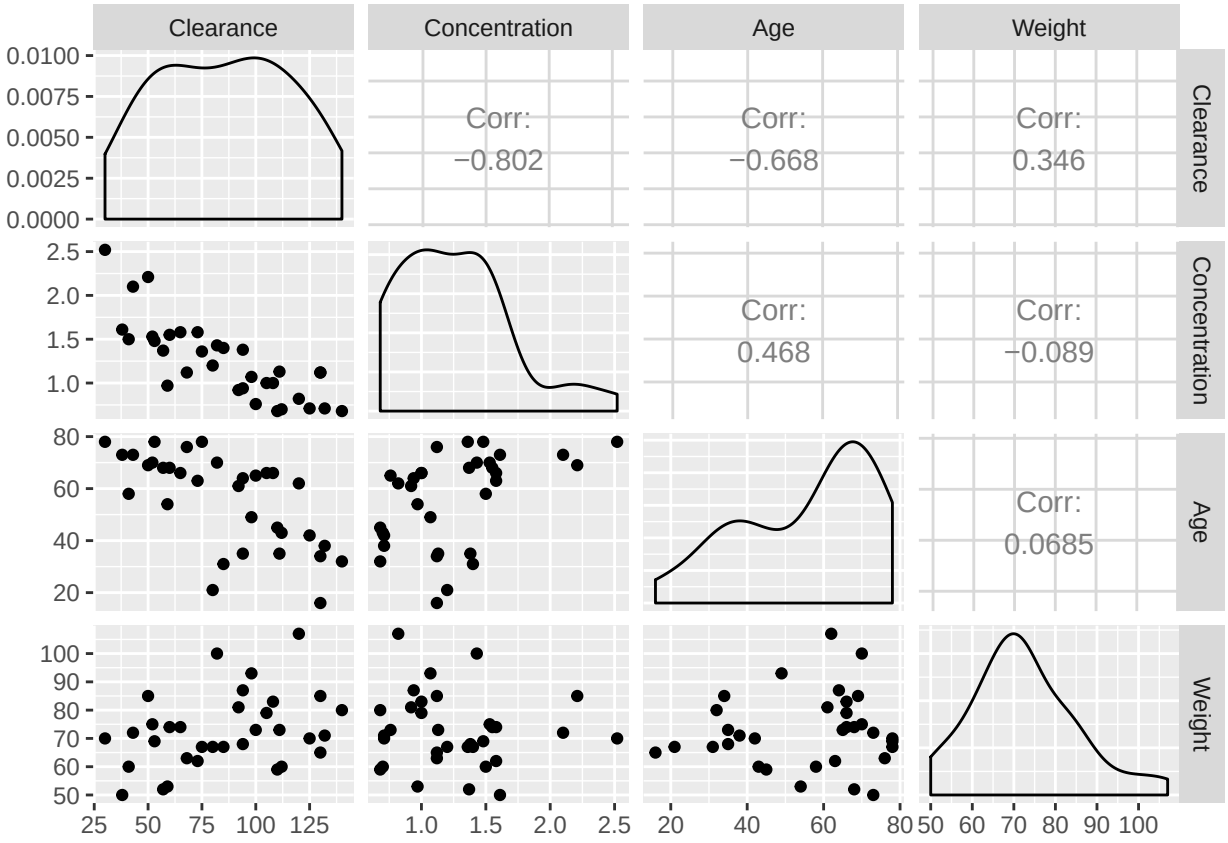
- Age is left-skewed. It appears that there were many more subjects between 60-80 than other age bands.
- Concentration is right-skewed. There were significantly more "small" doses than larger ones which is expected.

- Weight appears slightly right-skewed but may be normal.

b

Obtain the scatterplot matrix. Also obtain the correlation matrix of the X variables. What do the scatter plots suggest about the nature of the functional relationship between the response variable Y and each predictor variable? Discuss. Are any serious multicollinearity problems evident? Explain.

```
ggpairs(kidney)
```



Clearance and Concentration have the most significant correlation at -0.802 followed by Clearance and Age at -0.668. The other variables do not have any noteworthy correlations thus there is no clear relationship between a variable and the predictors. There may be weak multicollinearity between Concentration and Age as the correlation coefficient for Age and Clearance so that should be accounted for in the analysis of this data.

c

Fit the multiple regression function containing the three predictor variables as first-order terms. Does it appear that all predictor variables should be retained?

```
kidney.m.full <- lm(Clearance ~ ., data = kidney)
kidney.m.full %>% tidy %>% kable
```

term	estimate	std.error	statistic	p.value
(Intercept)	120.0472828	14.7737047	8.125740	0.00e+00
Concentration	-39.9393269	5.5999539	-7.132081	1.00e-07
Age	-0.7367673	0.1413940	-5.210738	1.41e-05
Weight	0.7764186	0.1718849	4.517085	9.69e-05

There is convincing evidence that each of the predictors are significant in the presence of others so that they should be retained. Typically multicollinearity causes larger standard errors and thus larger p-values but the p-values here are small so the relationship between Concentration and Age do not appear to have a significant effect.

9.16

a

> Using first-order and second-order terms for each of the three predictor variables (center around the mean) in the pool of potential X variables (including cross products of the first order terms), find the three best hierarchical subset regression models according to the C_p criterion.

```
kidney.m.full2 <- lm(Clearance ~ poly(Concentration, 2) + poly(Age, 2) + poly(Weight, 2) + (Concentrati
source('~/Dropbox/AppliedStats/RegressionAnalysis/fortify_leaps.R')
models <- leaps::regsubsets(Clearance ~ poly(Concentration, 2) + poly(Age, 2) + poly(Weight, 2) + (Conc
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,
## force.in = force.in, : 3 linear dependencies found
```

```
## Reordering variables and trying again:
```

```
models.fort <- fortify.regsubsets(models)
```

```
## Loading required package: plyr
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following object is masked from 'package:purrr':
##
##     compact
```

According to the C_p criterion, the following models are considered "best"

- $Y_i = \beta_0 + \beta_1 Weight^2$
- $Y_i = \beta_0 + \beta_1 Concentration^2$
- $Y_i = \beta_0 + \beta_1 Age^2$

In order to be viable, a model must contain its first order interactions as well. Filtering for those models, these three are the best:

- $Y_i = \beta_0 + \beta_1 Weight$
- $Y_i = \beta_0 + \beta_1 Age$
- $Y_i = \beta_0 + \beta_1 Concentration$

b

Is there much difference in C_p for the three best subset models?

Between the first three specified models, there is not much of a difference in C_p. Between the actual three that would be chosen, there is a significant difference.

- $Y_i = \beta_0 + \beta_1 Weight$ - 133.9048
- $Y_i = \beta_0 + \beta_1 Age$ - 73.50611
- $Y_i = \beta_0 + \beta_1 Concentration$ - 37.08015

9.19

a

Using the same pool of potential X variables as in Problem 9.16a, find the best subset of variables according to forward stepwise regression with α limits of 0.10 and 0.15 to add or delete a variable, respectively.

1. $Y_i = \beta_0 + \beta_1 Age + \beta_2 Concentration + \beta_3 Concentration \times Age + \beta_4 Weight^2$ - 165.74
2. $Y_i = \beta_0 + \beta_1 Age + \beta_2 Weight + \beta_3 Concentration + \beta_4 Concentration \times Age + \beta_5 Weight^2$ - 165.74*
3. $Y_i = \beta_0 + \beta_1 Age + \beta_2 Weight + \beta_3 Concentration + \beta_4 Concentration \times Age + \beta_4 Concentration \times Weight + \beta_5 Weight^2$

b

How does the best subset according to forward stepwise regression compare with the best subset according to the C_p criterion obtained in 9.16a?

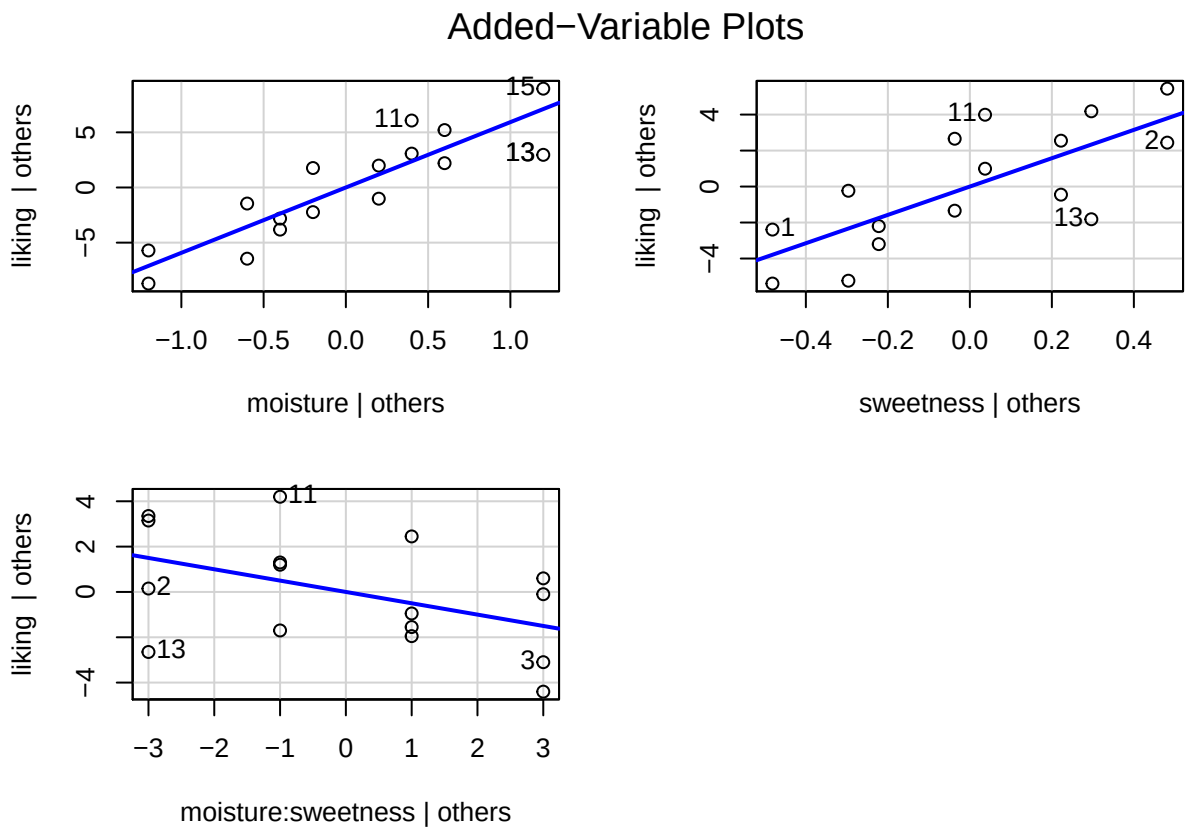
Stepwise selection only yielded one model that could be used among its top 3, (2). Coincidentally, this model has the lowest AIC so is the best fit. The models chosen by C_p were much simpler than that chosen by Stepwise selection aka AIC.

10.5

a

Prepare an added-variable plot for each of the predictor variables.

```
car::avPlots(brand.m.full)
```



b

Do your plots in part (a) suggest that the regression relationshups in the fitted regression function in 6.15c are inappropriate for any of the predictor variables? Explain.

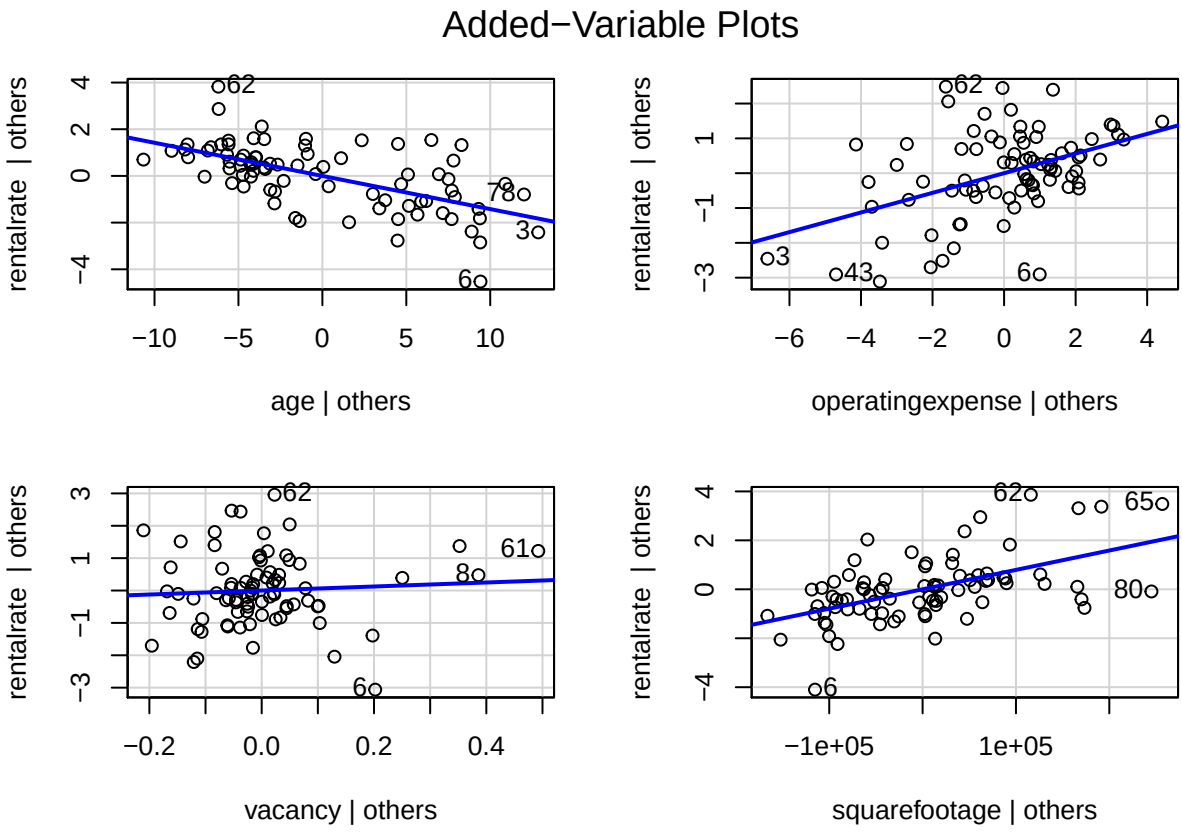
There is a distinctive pattern between liking given all the variables in the model and each of the predictors given that the other variables are present. This agrees with the interpretation of the regression coefficients provided by the results of two-tailed t-tests for each predictor.

10.8

a

Prepare an added-variable plot for each of the predictor variables.

```
properties <- read.csv("~/Downloads/commercialproperties.csv")
properties.m.full <- lm(rentalrate ~ ., data = properties)
car::avPlots(properties.m.full)
```



b

Do your plots in part (a) suggest that the regression relationshups in the fitted regression function in 6.18c are inappropriate for any of the predictor variables? Explain.

There are patterns for each of the predictors except the partial regression of vacancy on rental rate. This indicates that age, operating expense, and square footage significantly affect rentalrates whereas vacancy does not.

10.12

a

Obtain the studentized deleted residuals and identify any outling Y observations. Use the Bonferroni outlier test procedure with $\alpha = 0.01$. State the decision rule and conclusion.

```
augment(properties.m.full) %>%
  mutate(
    stud.res = rstudent(properties.m.full),
    is.outlier = stud.res > qt(1 - 0.01 / (2 * nrow(.)), nrow(.) - length(.) - 1)
  ) %>%
  filter(is.outlier)
```

```
## # A tibble: 0 x 14
```

```
## # ... with 14 variables: rentalrate <dbl>, age <int>,
## #   operatingexpense <dbl>, vacancy <dbl>, squarefootage <int>,
## #   .fitted <dbl>, .se.fit <dbl>, .resid <dbl>, .hat <dbl>, .sigma <dbl>,
## #   .cooksd <dbl>, .std.resid <dbl>, stud.res <dbl>, is.outlier <lgl>
```

There is no evidence that there are any outliers for the values with $\alpha = 0.01$ (bonferroni outlier test).

b

Obtain the diagonal elements of the hat matrix. Identify any outlying X observations.

```
# diagonal elements of H
H <- influence(properties.m.full)$hat
(H > 2 * mean(H))
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12
## FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
##    13    14    15    16    17    18    19    20    21    22    23    24
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    25    26    27    28    29    30    31    32    33    34    35    36
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    37    38    39    40    41    42    43    44    45    46    47    48
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    49    50    51    52    53    54    55    56    57    58    59    60
## FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    61    62    63    64    65    66    67    68    69    70    71    72
##  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    73    74    75    76    77    78    79    80    81
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
(H > 2 * mean(H)) %>% as_tibble %>% count %>% kable
```

```
## Warning: Calling `as_tibble()` on a vector is discouraged, because the behavior is likely to change i
## This warning is displayed once per session.
```

value	freq
FALSE	76
TRUE	5

There are 5 outliers according to the diagonal elements of the Hat matrix. Cases 3, 8, 53, 61, 65

c

The researcher wishes to estimate the rental rates of a property whose age is 10 years, operating expense sand taxes are 12, occupancy is 0.05, square footage is 350K. Use (10.29) to determine whether this estimate will involve a hidden extrapolation.

```
X <- properties %>%
  select(-rentalrate) %>% mutate(intercept = 1) %>%
  select(intercept, age, operatingexpense, vacancy, squarefootage) %>%
  as.matrix
X.new <- data.frame(intercept = 1, age = 10, operatingexpense = 12, vacancy = 0.05, squarefootage = 350

pred.lev <- X.new %*% solve(t(X) %*% X) %*% t(X.new)

augment(properties.m.full) %>%
```

```
  select(.hat) %>%
  summarise_all(c("min", "max", "median")) %>%
  kable
```

min	max	median
0.0241988	0.3036714	0.0481761

```
pred.lev
```

```
##             [,1]
## [1,] 0.05292296
```

The estimate's leverage (0.0529) falls within range of existing leverage in the models thus these values are not considered an extrapolation for this model.

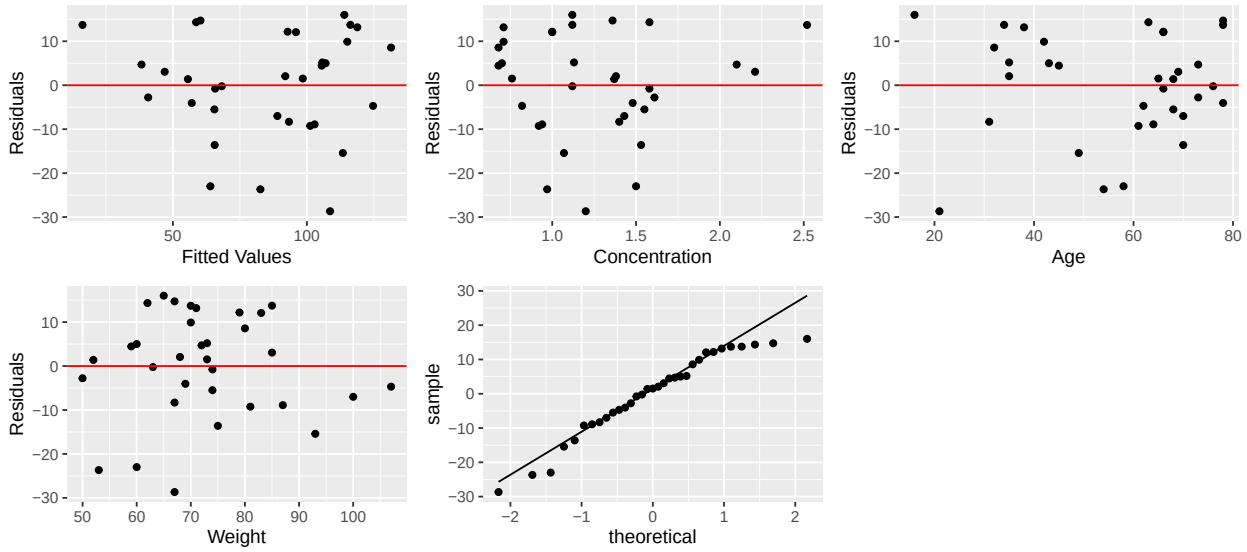
10.21

b

> Obtain the residuals and plot them separate against $\hat{Y}$ and each of the predictor variables. Also prepare a normal probability plot of the residuals.

```
kidney.p.resid.fit <-
  kidney.m.full %>%
  qplot(.fitted, .resid, data = .) +
  geom_hline(yintercept = 0, color = "red") +
  ylab("Residuals") +
  xlab("Fitted Values")

kidney.p.resid.concentration <-
  kidney.m.full %>%
  qplot(Concentration, .resid, data = .) +
  geom_hline(yintercept = 0, color = "red") +
  ylab("Residuals")

kidney.p.resid.age <-
  kidney.m.full %>%
  qplot(Age, .resid, data = .) +
  geom_hline(yintercept = 0, color = "red") +
  ylab("Residuals")

kidney.p.resid.weight <-
  kidney.m.full %>%
  qplot(Weight, .resid, data = .) +
  geom_hline(yintercept = 0, color = "red") +
  ylab("Residuals")

kidney.p.qq <-
  kidney.m.full %>%
  ggplot(aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line()

grid.arrange(
```

```
  kidney.p.resid.fit,
  kidney.p.resid.concentration,
  kidney.p.resid.age,
  kidney.p.resid.weight,
  kidney.p.qq,
  ncol = 3
)
```



Fitted vs Residuals

- no distinctive pattern
- spread seems consistent so constant variance seems likely
- there appear to be a few outliers with residual values > 20

Concentration vs Residuals

- no distinctive pattern
- consistent spread for smaller values of concentration
- three potential outliers appear that may be same values in the previous plot

Age vs Residuals

- no distinctive pattern
- even spread around y = 0 so constant variance is possible
- three potential outliers appear that may be same values in the previous plots

Weight vs Residuals

- no distinctive pattern
- even spread around y = 0 so constant variance is possible. There do appear to be a cluster of values for residuals > 0.
- three potential outliers appear that may be same values in the previous plots

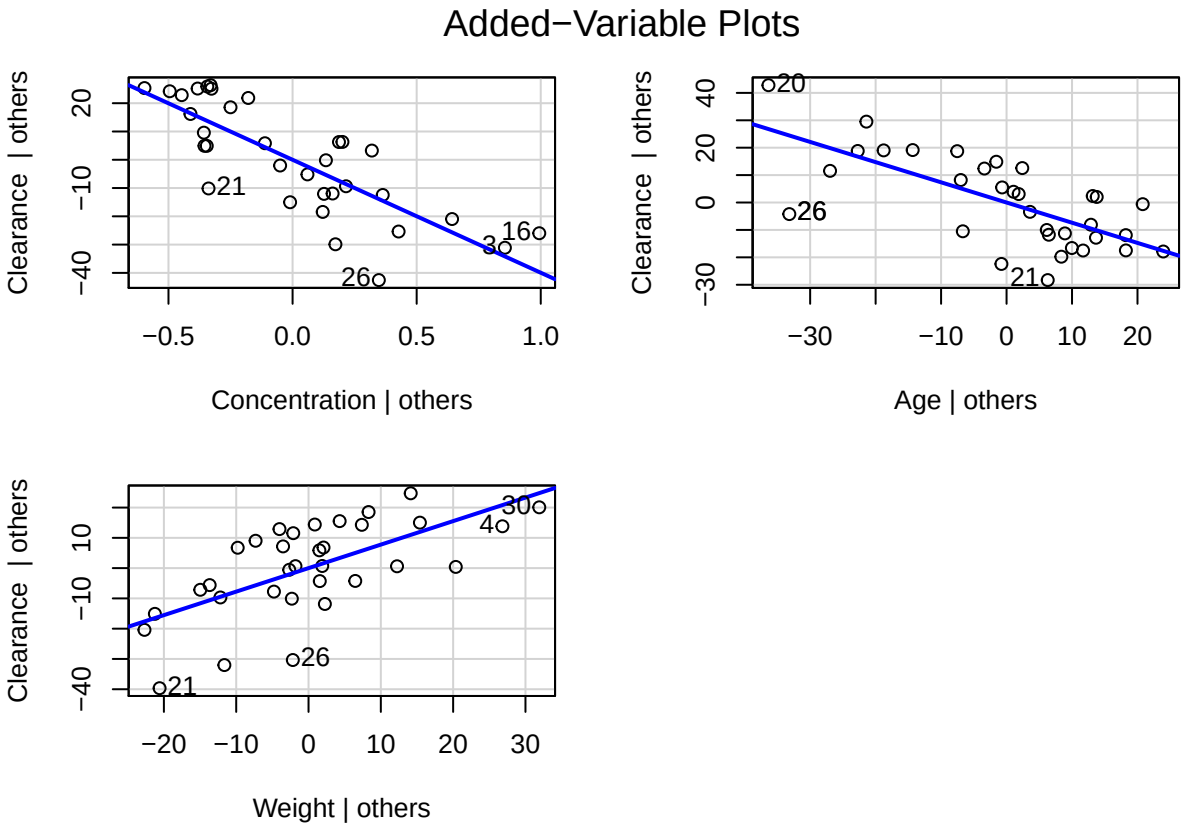
Normal probability plot

- The middle area of the graph is close to the line but the right-tailed values deviate from the graph indicating that extreme values may be non-normal.

c

Prepare separate added-variable plots.

```
car::avPlots(kidney.m.full)
```



d

Do plots in parts (b) and (c) suggest that the regression model should be modified?

```
shapiro.test(kidney.m.full$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  kidney.m.full$residuals
## W = 0.94004, p-value = 0.0681
```

Given the lack of patterns in the residual plots, a linear model appears appropriate for this data. The outlier values are concerning however so it would be worth while to fit the model with and without the outliers to assess their impact on the model. The normality assumption appears suspect and a follow-up Shapiro

Wilk test indicates that there is weak evidence that the data is non-normal (p-value = 0.0681). Use of that assumption should proceed with caution. It should not be an issue for this dataset however given that the t-test (and thus one of the statistics being used to determine effectiveness of parameters) is robust to assumptions of non-normality with a large enough sample size.

In conclusion, the regression model appears to be adequate and no further modifications are necessary.