STAT 151A Optional Problems (M2): Solutions

Billy Fang

These are rough sketches for the solutions. Some computational steps are omitted for brevity.

1

(a) False. The $2 \cdot (\text{number of parameters})$ term discourages choosing models with larger number of parameters.

(b) False. See end of "Variable selection 1" lecture notes.

(c) False. Although $\frac{\widehat{e}_i}{\sigma\sqrt{1-h_i}}$ has variance equal to 1, the studentized residual (a.k.a. standardized residual) $r_i = \frac{\widehat{e}_i}{\widehat{\sigma}\sqrt{1-h_i}}$ will not have variance 1 in general, due to replacing the deterministic quantity σ by the random variable $\widehat{\sigma}$. Note that the standardized predicted residual t_i also does not have variance one, since it follows a t distribution.

(d) False. Cook's distance detects influential points. In particular, it can detect points with high residuals (outlier in y) but low leverage (not outlier in x). Note also that it can also detect high leverage points that are not outliers [in y].

(e) False. It is possible that the fit does not change after removing a high leverage point.

2

2.1 Exercise 11.1

In simple regression,

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}.$$

In part of a question on Homework 2, you showed that

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{\sum_j (X_j - \overline{X})^2} \begin{bmatrix} \frac{1}{n}\sum_j X_j^2 & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix}.$$

Thus with $\overline{x}_i := \begin{bmatrix} 1 \\ X_i \end{bmatrix}$, we have

$$\begin{aligned} h_i &= \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i \\ &= \frac{1}{\sum_j (X_j - \overline{X})^2} \begin{bmatrix} 1 & X_i \end{bmatrix} \begin{bmatrix} \frac{1}{n}\sum_j X_j^2 & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix} \begin{bmatrix} 1 \\ X_i \end{bmatrix} \\ &= \frac{1}{\sum_j (X_j - \overline{X})^2} \left(\frac{1}{n} \sum_j X_j^2 - 2\overline{X} X_i + X_i^2 \right) \\ &= \frac{1}{\sum_j (X_j - \overline{X})^2} \left(\frac{1}{n} \sum_j X_j^2 - \overline{X}^2 + (X_i - \overline{X})^2 \right) \\ &= \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_j (X_j - \overline{X})^2}, \end{aligned}$$

where again we use the fact that $\frac{1}{n}\sum_j (X_j - \overline{X})^2 = \frac{1}{n}\sum_j X_j^2 - \overline{X}^2$ (see HW2 solutions).

2.2 Exercise 11.3

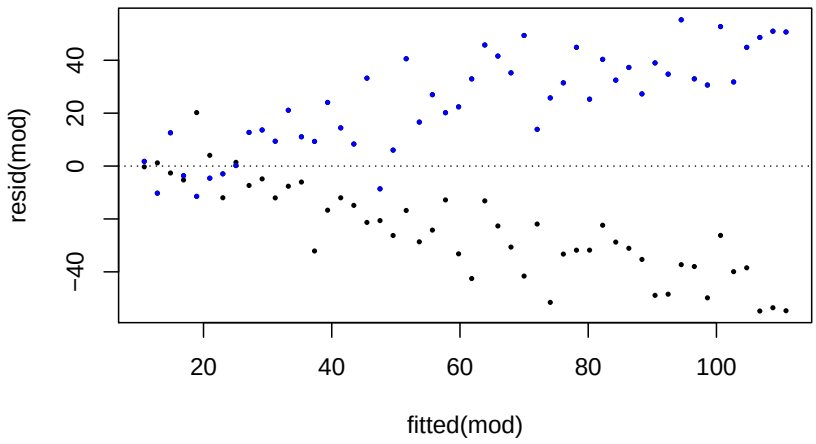
In one-way ANOVA, $X^\top X$ is a diagonal matrix with diagonal entries $n_1, \ldots, n_m$, where m is the number of groups and n_j is the number of subjects in group j. If a subject in group j is represented by a vector x_i with all zeros except a one in the jth entry ($x_i^\top$ row of the design matrix X), we have

$$h_i = x_i^\top (X^\top X)^{-1} x_i = \frac{1}{n_j}.$$

If the group sizes are the same, then the hat values are the same.

To see why this generalizes to higher-way ANOVA models with balanced design (equal cell means), note that the design matrix for higher-way ANOVA can be set up to look like a one-way ANOVA with a level for each cell.

2.3 Exercise 12.2

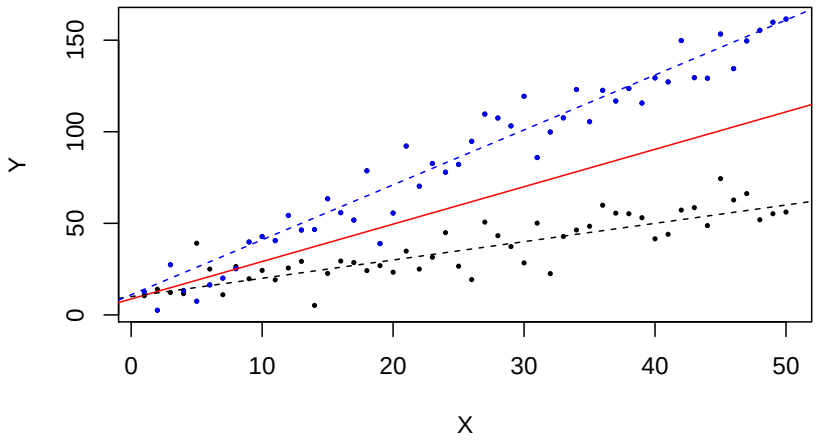


The residuals seem to have increasing variance. If we highlight the points corresponding to $D = 1$ in blue, it is clear that the hidden group structure is responsible for this.

From the model, you can see that there are two sub-models corresponding to $D = 0$ and $D = 1$ respectively.

$$Y = 10 + X + \epsilon, \qquad (D = 0),$$
$$Y = 11 + 3X + \epsilon, \qquad (D = 1),$$

shown as dotted lines in the plot below. Thus the increasing variance in the residuals from regressing Y on X is simply due to the increasing gap between the above two sub-models.



Code:

```
n <- 100
X <- rep(1:50, 2)
D <- c(rep(0, 50), rep(1, 50))
Y <- 10 + X + D + 2 * X * D + rnorm(n, sd=10)
mod <- lm(Y ~ X)

pdf("resid.pdf", width=6, height=4)
plot(fitted(mod), resid(mod), pch=20, cex=0.5)
points(fitted(mod)[51:100], resid(mod)[51:100], pch=20, cex=0.5, col='blue')
abline(h=0, lty=3)
dev.off()

pdf("plot.pdf", width=6, height=4)
plot(X, Y, pch=20, cex=0.5)
points(X[51:100], Y[51:100], pch=20, cex=0.5, col='blue')
coefs <- coef(mod)
abline(a=coefs[1], b=coefs[2], col='red')
abline(a=10, b=1, lty=2)
abline(a=11, b=3, lty=2, col='blue')
dev.off()
```

2.4 Exercise 12.3

For brevity I will use the notation σ^2 in place of σ_ϵ^2. There is a typo in part (a); the Σ in the exponential should be Σ^{-1}. There is a typo in part (b); the E_i/w_i should be $w_i E_i$.

(a) **Approach 1.** Note $y_i \sim N(x_i^\top \beta, \sigma^2/w_i^2)$. Because the ϵ_i are independent of each other, the y_i are independent

of each other. The likelihood of the data is

$$\begin{aligned} L(\beta, \sigma^2) &= p(y; \beta, \sigma^2) \\ &= \prod_{i=1}^{n} p(y_i; \beta, \sigma^2) && \text{independence} \\ &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2/w_i^2}} \exp\left(-\frac{(y_i - x_i^\top \beta)^2}{2\sigma^2/w_i^2} \right) \right] \\ &= \frac{1}{(2\pi)^{n/2} \sqrt{\prod_{i=1}^{n} (\sigma^2/w_i^2)}} \exp\left[-\frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - x_i^\top \beta)^2}{\sigma^2/w_i^2} \right] \end{aligned}$$

To conclude, check that with $\Sigma = \sigma^2 \operatorname{diag}(w_1^{-2}, \ldots, w_n^{-2})$ as defined in the problem, we have the following two equalities

$$|\Sigma| = \prod_{i=1}^{n} \frac{\sigma^2}{w_i^2},$$
$$(y - X\beta)^\top \Sigma^{-1} (y - X\beta) = \sum_{i=1}^{n} \frac{(y_i - x_i^\top \beta)^2}{\sigma^2/w_i^2}$$

and substitute them into the expression above.

Approach 2: multivariate normal. Note that the random vector ϵ follows the multivariate normal distribution with mean zero [vector] and covariance matrix $\Sigma = \sigma^2 \operatorname{diag}(w_1^{-2}, \ldots, w_n^{-2})$. Thus, y also follows a multivariate normal distribution, namely

$$y \sim N(X\beta, \Sigma).$$

Then the likelihood is simply the density of the this multivariate normal distribution, evaluated at y.

(b) By taking the logarithm of part (a) and applying the definition $\Sigma = \sigma^2 W^{-1}$ given in the problem, we see that the log likelihood is

$$-\frac{n}{2} \log(2\pi) - \frac{1}{2} \log|\sigma^2 W^{-1}| - \frac{1}{2\sigma^2} (y - X\beta)^\top W (y - X\beta).$$

The gradient with respect to β is

$$\frac{1}{\sigma^2} X^\top W (y - X\beta).$$

Setting this equal to zero and solving for β yields the expression for $\widehat{\beta}$.

Now, we plug in $\widehat{\beta}$ into the log likelihood and compute the derivative with respect to σ^2. Noting that $\log|\sigma^2 W^{-1}| = n \log(\sigma^2) + \log|W^{-1}|$, the derivative of the log likelihood is

$$-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} (y - X\widehat{\beta})^\top W (y - X\widehat{\beta}).$$

Setting this equal to zero, solving for σ^2, and using the substitution $E = y - X\widehat{\beta}$ yields

$$\widehat{\sigma}^2 = \frac{1}{n} E^\top W E = \frac{\sum_{i=1}^{n} (w_i E_i)^2}{n}.$$

(c) Choosing β to minimize the weighted sum of squares $\sum_{i=1}^{n} w_i^2 (y - X\beta)^2 = (y - X\beta)^\top W (y - X\beta)$ is the same as maximizing the likelihood.

(d) The covariance matrix of $\widehat{\beta}$ is

$$\begin{aligned}\operatorname{Cov}(\widehat{\beta}) &= \operatorname{Cov}((X^\top W X)^{-1} X^\top W y) \\ &= (X^\top W X)^{-1} X^\top W \operatorname{Cov}(y) W^\top X (X^\top W X)^{-1} \\ &= (X^\top W X)^{-1} X^\top W (\sigma^2 W^{-1}) W^\top X (X^\top W X)^{-1} \\ &= \sigma^2 (X^\top W X)^{-1}.\end{aligned}$$

Replacing σ^2 with the MLE estimate yields the result.

2.5 Exercise 12.4

Imitate the proof of Gauss-Markov in the textbook.

Let $M = (X^\top W X)^{-1} X^\top W$ so that $\widehat{\beta} = My$ is the above estimator. Let $\widetilde{\beta} = (M + A)y$ be the BLUE estimator, where A is some matrix. Our goal is to show $A = 0$.

Because $\widetilde{\beta}$ is unbiased, we have

$$\begin{aligned}\beta = \mathbb{E}\widetilde{\beta} &= (M + A)\mathbb{E}y = (M + A)X\beta \\ &= (X^\top W X)^{-1} X^\top W X \beta + AX\beta \\ &= \beta + AX\beta,\end{aligned}$$

so $AX\beta = 0$ for any β, and thus $AX = 0$.

The covariance matrix of $\widetilde{\beta}$ is

$$\begin{aligned}\operatorname{Cov}(\widetilde{\beta}) &= (M + A)\operatorname{Cov}(y)(M + A)^\top \\ &= \sigma^2 (M + A) W^{-1} (M + A)^\top \\ &= \sigma^2 (M W^{-1} M^\top + \underbrace{M W^{-1} A^\top}_{=0} + \underbrace{A W^{-1} M^\top}_{=0} + A W^{-1} A^\top) \\ &= \sigma^2 (M W^{-1} M^\top + A W^{-1} A^\top),\end{aligned}$$

where the last step uses $AX = 0$ to deduce $M W^{-1} A^\top = (X^\top W X)^{-1} X^\top W W^{-1} A^\top = (X^\top W X)^{-1} (AX)^\top = 0$.

The diagonal entries are

$$\operatorname{Var}(\widetilde{\beta}_j) = \sigma^2 \left(\sum_{i=1}^n \frac{m_{ji}^2}{w_i^2} + \sum_{i=1}^n \frac{a_{ji}^2}{w_i^2} \right).$$

Setting all the entries of A to zero simultaneously minimizes $\operatorname{Var}(\widetilde{\beta}_j)$ for all j.

2.6 Exercise 12.5

(a) We already know a formula for B. Thus,

$$\begin{aligned}\operatorname{Var}(B) &= \operatorname{Var}\left(\frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_j (X_j - \overline{X})^2} \right) \\ &= \frac{1}{\left[\sum_j (X_j - \overline{X})^2\right]^2} \operatorname{Var}\left(\sum_i (X_i - \overline{X}) Y_i \right) \qquad \text{note } \overline{Y} \sum_i (X_i - \overline{X}) = 0 \\ &= \frac{\sum_i (X_i - \overline{X})^2 \sigma_i^2}{\left[\sum_j (X_j - \overline{X})^2\right]^2}\end{aligned}$$

Let

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}.$$

Then

$$(\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1} = \begin{bmatrix} \sum_i w_i^2 & \sum_i w_i^2 X_i \\ \sum_i w_i^2 X_i & \sum_i w_i^2 X_i^2 \end{bmatrix}^{-1} = \frac{1}{(\sum_i w_i^2)(\sum_i w_i^2 X_i^2) - (\sum_i w_i^2 X_i)^2} \begin{bmatrix} \sum_i w_i^2 X_i^2 & -\sum_i w_i^2 X_i \\ -\sum_i w_i^2 X_i & \sum_i w_i^2 \end{bmatrix}.$$

So, the weighted least squares estimator is

$$\begin{aligned} &(\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W} y \\ &= (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1} \begin{bmatrix} \sum_i w_i^2 Y_i \\ \sum_i w_i^2 X_i Y_i \end{bmatrix} \\ &= \frac{1}{(\sum_i w_i^2)(\sum_i w_i^2 X_i^2) - (\sum_i w_i^2 X_i)^2} \begin{bmatrix} (\sum_i w_i^2 X_i^2)(\sum_i w_i^2 Y_i) - (\sum_i w_i^2 X_i)(\sum_i w_i^2 X_i Y_i) \\ (\sum_i w_i^2)(\sum_i w_i^2 X_i Y_i) - (\sum_i w_i^2 X_i)(\sum_i w_i^2 Y_i). \end{bmatrix} \end{aligned}$$

The slope $\widehat{\beta}$ is the second element of this vector.

Note that with $\widetilde{X}$ as defined in the problem, we have

$$\sum_i w_i^2 (X_i - \widetilde{X})^2 = \sum_i w_i^2 X_i^2 - 2\widetilde{X} \sum_i w_i^2 X_i + \widetilde{X}^2 \sum_i w_i^2 = \sum_i w_i^2 X_i^2 - \frac{(\sum_i w_i^2 X_i)^2}{\sum_i w_i^2},$$

so the second element of the above vector can be written as

$$\widehat{\beta} = \frac{\sum_i w_i^2 X_i Y_i - \frac{(\sum_i w_i^2 X_i)(\sum_i w_i^2 Y_i)}{\sum_i w_i^2}}{\sum_i w_i^2 (X_i - \widetilde{X})^2}.$$

The variance is then

$$\begin{aligned} \mathrm{Var}(\widehat{\beta}) &= \frac{1}{(\sum_i w_i^2 (X_i - \widetilde{X})^2)^2} \mathrm{Var}\left(\sum_i w_i^2 X_i Y_i - \frac{(\sum_i w_i^2 X_i)(\sum_i w_i^2 Y_i)}{\sum_i w_i^2} \right) \\ &= \frac{1}{(\sum_i w_i^2 (X_i - \widetilde{X})^2)^2} \mathrm{Var}\left(\sum_i \left(X_i - \frac{\sum_j w_j^2 X_i}{\sum_j w_j^2} \right) w_i^2 Y_i \right) \\ &= \frac{1}{(\sum_i w_i^2 (X_i - \widetilde{X})^2)^2} \mathrm{Var}\left(\sum_i w_i^2 (X_i - \widetilde{X}) Y_i \right) \\ &= \frac{\sigma^2}{(\sum_i w_i^2 (X_i - \widetilde{X})^2)^2} \sum_i w_i^4 (X_i - \widetilde{X})^2 X_i^2 \\ &= \frac{\sigma^2}{\sum_i w_i^2 (X_i - \widetilde{X})^2}, \end{aligned}$$

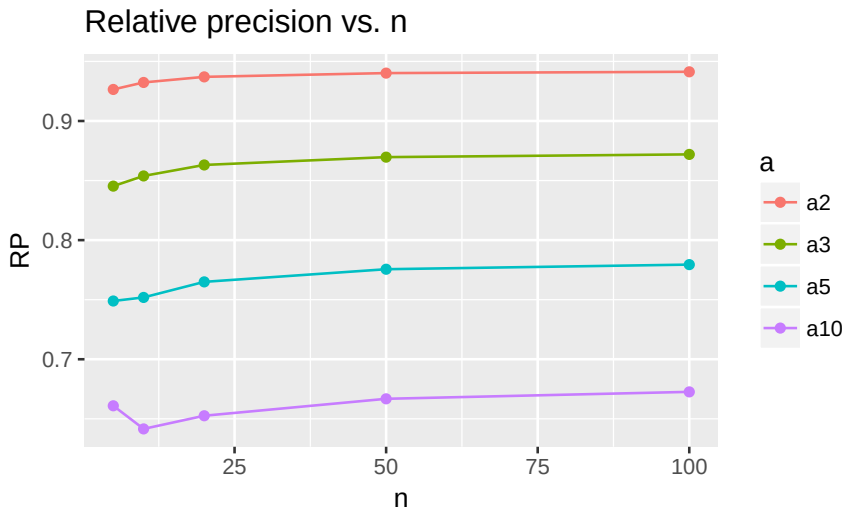
where in the last step we note that $w_i^2 = X_i^{-2}$, since $\sigma_i^2 = \sigma^2 X_i^2 = \sigma^2 / w_i^2$.

(b)

$$\frac{\mathrm{SD}(\widehat{\beta})}{\mathrm{SD}(B)} = \sqrt{\frac{\mathrm{Var}(\widehat{\beta})}{\mathrm{Var}(B)}} = \frac{\sigma / \sqrt{\sum_i w_i^2 (X_i - \widetilde{X})^2}}{\sigma \sqrt{\sum_i X_i^2 (X_i - \overline{X})^2} / \sum_i (X_i - \overline{X})^2} = \frac{\sum_i (X_i - \overline{X})^2}{\sqrt{(\sum_i X_i^2 (X_i - \overline{X})^2)(\sum_i X_i^{-2} (X_i - \widetilde{X})^2)}},$$

where $\widetilde{X} = \frac{\sum_i w_i^2 X_i}{\sum_i w_i^2} = \frac{\sum_i X_i^{-1}}{\sum_i X_i^{-2}}$.

From the plot below, we see that OLS becomes much less precise than WLS the smaller n is and the larger a is.



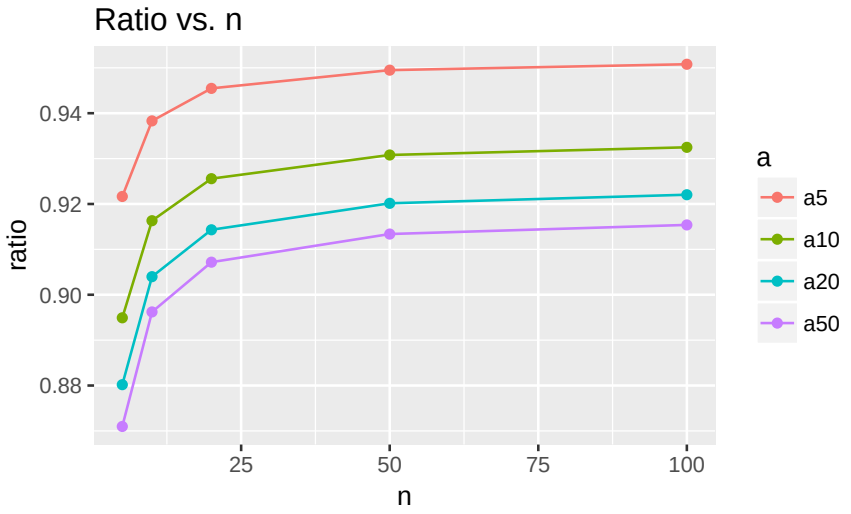
Code:

```
library(ggplot2)
library(reshape2)

RelPrec <- function(x) {
  # compute directly
  VB <- sum((x - mean(x))^2 * x^2) / sum((x - mean(x))^2)^2
  xtilde <- sum(1/x) / sum(1/x^2)
  Vbeta <- 1 / sum(1 / x^2 * (x - xtilde)^2)
  return(sqrt(Vbeta / VB))
}

RelPrec2 <- function(x) {
  # compute using simplified formula
  # should be same as above
  xtilde <- sum(1/x) / sum(1/x^2)
  return(
    sum((x - mean(x))^2) / sqrt(
      sum(x^2 * (x - mean(x))^2) * sum(1 / x^2 * (x - xtilde)^2)
    )
  )
}

a.vec <- c(2,3,5,10)
n.vec <- c(5,10,20,50,100)

RP <- matrix(0, length(n.vec), length(a.vec))

for (i in 1:length(a.vec)) {
  for (j in 1:length(n.vec)) {
    RP[j,i] <- RelPrec(seq(1, a.vec[i], length.out=n.vec[j]))
  }
}

RP.dat <- data.frame(cbind(n.vec, RP))
colnames(RP.dat) <- c("n", paste0("a", as.character(a.vec)))
```

```
RP.melt <- melt(RP.dat, id.vars="n", value.name="RP", variable.name="a")
pdf("RP.pdf", width=5, height=3)
ggplot(RP.melt, aes(x=n, y=RP, color=a)) + geom_point() +
  geom_line(aes(group=a)) + ggtitle("Relative precision vs. n")
dev.off()
```

(c) Note $\overline{\sigma}^2 = \frac{\sigma^2}{n} \sum_i X_i^2$.

The plot below shows that the usual variance estimate underestimates the actual variance more and more the larger a is and the smaller n is.



Code:

```
library(ggplot2)
library(reshape2)


Rat <- function(x, n) {
  sigmabarsq <- sum(x^2) / n
  Vhat <- sigmabarsq / sum((x - mean(x))^2) -
    sum((x - mean(x))^2 * (x^2 - sigmabarsq)) / (n-2) / sum((x - mean(x))^2)^2
  V <- sum((x - mean(x))^2 * x^2) / sum((x - mean(x))^2)^2
  return(sqrt(Vhat / V))
}

a.vec <- c(5, 10, 20, 50)
n.vec <- c(5, 10, 20, 50, 100)

rat <- matrix(0, length(n.vec), length(a.vec))

for (i in 1:length(a.vec)) {
  for (j in 1:length(n.vec)) {
    rat[j,i] <- Rat(seq(1, a.vec[i], length.out=n.vec[j]), n.vec[j])
  }
}

rat.dat <- data.frame(cbind(n.vec, rat))
```

```
colnames(rat.dat) = c("n", paste0("a", as.character(a.vec)))
rat.melt <- melt(rat.dat, id.vars="n", variable.name="a", value.name="ratio")
pdf("ratio.pdf", width=5, height=3)
ggplot(rat.melt, aes(x=n, y=ratio, color=a)) + geom_point() +
  geom_line(aes(group=a)) + ggtitle("Ratio vs. n")
dev.off()
```

3

We need to check that $h(y) := \arcsin(\sqrt{y})$ satisfies

$$h'(\mathbb{E}Y) \propto \frac{1}{\sqrt{\mathrm{Var}(Y)}}.$$

Note $\mathbb{E}[Y_i] = \mathbb{E}[X_i]/m = p_i$ and $\mathrm{Var}(Y_i) = \mathrm{Var}(X_i)/m^2 = p_i(1-p_i)/m = \mathbb{E}[Y_i](1-\mathbb{E}[Y_i])/m$. Thus we want

$$h'(y) \propto \sqrt{\frac{m}{y(1-y)}}.$$

Indeed, the chain rule implies

$$h'(y) = \frac{1}{\sqrt{1-y}} \cdot \frac{1}{2\sqrt{y}}.$$

4

(a)
$$\mathrm{tr}(H) = \mathrm{tr}(X(X^\top X)^{-1}X^\top) = \mathrm{tr}((X^\top X)^{-1}X^\top X) = \mathrm{tr}(I_{p+1}) = p+1.$$

(b) Since hat values are nonnegative and sum to a constant $p+1$, a large leverage value will necessarily force the other leverage values to be small.

5 Proving the slope and residuals from simple regression on CR plot are the same as coefficient and residuals from multiple regression

Approach 1. Let x_j be the [column] vector with values $x_{1j}, \ldots, x_{nj}$, and let $\widehat{e}^{(j)}$ be the [column] vector of partial residuals $\widehat{e}_1^{(j)}, \ldots, \widehat{e}_n^{(j)}$. The vector of fitted values from the simple regression of $\widehat{e}^{(j)}$ on x_j is the projection of $\widehat{e}^{(j)}$ onto $\mathrm{span}(\vec{1}, x_j)$. We already know $\widehat{e}$ is perpendicular to this span, since it is perpendicular to the column space of the entire design matrix. Thus, if we let $\widetilde{H}$ be the projection onto $\mathrm{span}(\vec{1}, x_j)$, we have $\widetilde{H}\widehat{e} = 0$ and $\widetilde{H}x_j = x_j$, and thus

$$\widetilde{H}\widehat{e}^{(j)} = \widetilde{H}(\widehat{e} + \widehat{\beta}_j x_j) = 0 \cdot \vec{1} + \widehat{\beta}_j x_j.$$

Thus the slope from this simple regression is indeed $\widehat{\beta}_j$ from the full regression. Moreover, the residuals are precisely $\widehat{e}^{(j)} - \widehat{\beta}_j x_j = \widehat{e}$, the residuals from the full regression.

Approach 2. Let $\overline{x}_j = \frac{1}{n}\sum_{i=1}^n x_{ij}$. We define $\overline{\widehat{e}^{(j)}} := \frac{1}{n}\sum_i \widehat{e}_i^{(j)}$. Plugging in $\widehat{e}_i^{(j)} = \widehat{e}_i + \widehat{\beta}_j x_j$ and recalling that the sum of residuals $\widehat{e}_i$ is zero yields $\overline{\widehat{e}^{(j)}} = \widehat{\beta}_j \overline{x}_j$.

The slope of the simple regression can be computed explicitly from the formula for simple regression. Applying the above facts shows that the slope is

$$\frac{\sum_i (x_{ij} - \overline{x}_j)(\widehat{e}_i^{(j)} - \overline{\widehat{e}^{(j)}})}{\sum_i (x_{ij} - \overline{x}_j)^2} = \frac{\sum_i (x_{ij} - \overline{x}_j)(\widehat{e}_i^{(j)} - \widehat{\beta}_j \overline{x}_j)}{\sum_i (x_{ij} - \overline{x}_j)^2} = \frac{\sum_i (x_{ij} - \overline{x}_j)(\widehat{e}_i + \widehat{\beta}_j (x_{ij} - \overline{x}_j))}{\sum_i (x_{ij} - \overline{x}_j)^2}.$$

The vector $x_j - \overline{x}_j$ lies in the column space of X, so $\sum_i (x_{ij} - \overline{x}_j)\widehat{e}_i = (x_j - \overline{x}_j)^\top \widehat{e} = 0$ since the residual is orthogonal to the column space of X. Applying this to the above expression shows that the slope is simply $\widehat{\beta}_j$. Consequently, the residuals of the simple regression are $\widehat{e}^{(j)} - \widehat{\beta}_j x_j = \widehat{e}$, the residuals from the full regression.

6 Exercise 22.2

The formula for the F-statistic is

$$F_j = \frac{(\text{RSS}_j - \text{RSS})/(k+1-s_j)}{S_E^2},$$

so

$$(k+1-s_j)(F_j-1)+s_j = \frac{\text{RSS}_j - \text{RSS}}{S_E^2} - k - 1 + 2s_j = \frac{\text{RSS}_j}{S_E^2} + 2s_j - n,$$

where we used $S_E^2 = \text{RSS}/(n-k-1)$.

7 Exercise 22.3

Let one model have residual sum of squares RSS and s parameters, while a second model has RSS$'$ and s'. Suppose RSS$' = 3$RSS and $n - s' = 2(n-s)$. Then

$$\frac{\text{RSS}'}{n-s'} = 1.5 \cdot \frac{\text{RSS}}{n-s},$$

so adjusted R^2 prefers the first model, but

$$\frac{\text{RSS}'}{(n-s')^2} = 0.75 \cdot \frac{\text{RSS}}{(n-s)^2}$$

so GCV prefers the second model.