BIOSTAT/STAT 570: Key for Coursework 1

To be submitted to the course canvas site by 1:30pm Monday 11th October, 2021.

The data we analyze were collected on $n = 97$ men before radical prostatectomony and we take as response the log of prostate specific antigen (PSA) which was being forwarded in the paper as a preoperative marker (to predict the clinical stage of cancer). We will consider two covariates for modeling log PSA (`lpsa`) on patient i: log(weight) (`lweight`, x_{i1}) and seminal vesicle invasion (`svi`, $x_{i2} = 0/1$) and their interaction $x_{i3} = x_{i1} \times x_{i2}$. All logs are base e.

We will examine the model for PSA (ng/ml) with main effects due to log(prostate weight) (prostate weight in gm) and SVI, and their interaction.

We let Y_i represent log PSA, and (x_{i1}, x_{i2}, x_{i3}), the covariates, for individual i, $i = 1, \dots, n = 97$. We fit the model

$$y_i = \beta_0 + \sum_{j=1}^{3} x_{ij}\beta_j + \epsilon_i,$$

$i = 1, \dots, n$ in R, using least squares, and the output below (which has been edited slightly) was produced.

The computation part

1. Using R, reproduce every number in the handout using matrix and arithmetic operations.

The following R Code reproduces each number in the table at the bottom of the assignment.

```
library(lasso2)
data(Prostate)
attach(Prostate)
y <- Prostate$lpsa
lweight <- Prostate$lweight
svi <- Prostate$svi

### Calculations.
X <- cbind(1,svi,lweight,svi*lweight)
coeffs <- solve(t(X)%*%X)%*%t(X)%*%y
yhats <- X%*%coeffs
resids <- y-yhats
SSE <- sum((resids)^2)
```

```
varY <- SSE/(NROW(X)-NCOL(X))
StdEs <- sqrt(diag(solve(t(X)%*%X)*varY))
ts <- coeffs/StdEs
pvals <- 2*(1-pt(abs(ts), df=NROW(X)-NCOL(X)))

#### Format table to match handout.
results <- cbind(coeffs, StdEs,ts,pvals)
rownames(results) <- c("(Intercept)", "svi", "lweight", "svi:lweight")
colnames(results) <- c("Estimate", "Std. Error", "t-value", "p-value")
round(results,4)
cat("Residual Standard Error: " , round(sqrt(varY),4), "on", NROW(X)-NCOL(X),
"degrees of freedom")
```

The values in the table should be:

```
                Estimate Std. Error t-value p-value
(Intercept)  -0.5965     0.7127 -0.8370  0.4047
svi           4.1826     2.4149  1.7320  0.0866
lweight       0.7541     0.1946  3.8756  0.0002
svi:lweight  -0.7197     0.6425 -1.1202  0.2655

Residual Standard Error:  0.8969 on 93 degrees of freedom
```

The interpretation part: imagine this part will be read by a non-statistician

1. Based on the fitted model, provide an informative plot that summarizes the association between log PSA, log weight and SVI.

   There are many possible options for plots. Here is an example:

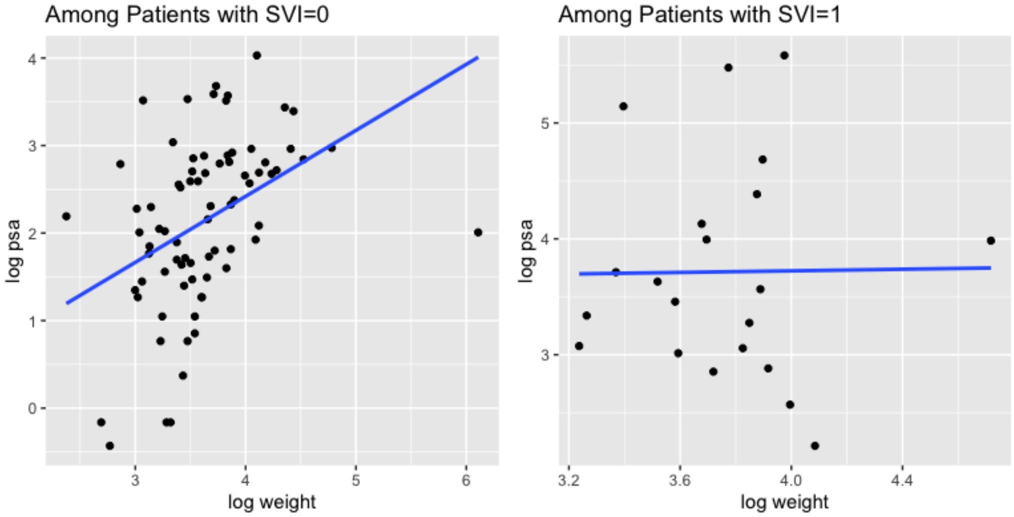


Figure 1: See appendix for code to reproduce this plot.

We see a positive association between log-weight and log-PSA among those with 0 *svi, as expected from our interpretation of* β_2 *below. On the other hand, we see essentially no association between log-PSA and log-sva among those with* $svi = 1$*, which also makes sense when we interpret the coefficients below. Plotting is also useful because we see the influence of one outlier with a very large weight.*

2. Give interpretations of each of the parameters β_j, $0 = 1, \ldots, 3$.

- $\hat{\beta}_0$: *The average log PSA for a patient whose log-weight is* 0 *(meaning that their weight is* 1 *gram) and whose seminal vesicle invasion is* 0*.*

$$\beta_0 = E\left[log(PSA) \mid lweight = 0, SVI = 0\right]$$

- $\hat{\beta}_1$*: The difference in average log PSA for two individuals with* $lweight = 0$ *(meaning their weight is* 1 *gram) if one individual has* $svi = 0$ *and the other has* $svi = 1$*.*

$$\beta_2 = E\left[log(PSA) \mid lweight = 0, SVI = 1\right] - E\left[log(PSA) \mid lweight = 0, SVI = 0\right]$$

- $\hat{\beta}_2$*: The change in average log PSA for a 1-unit increase in log weight for those with* $svi = 0$*.*

$$\beta_2 = E\left[log(PSA) \mid lweight, SVI = 0\right] - E\left[log(PSA) \mid lweight - 1, SVI = 0\right]$$

- $\hat{\beta}_3$: *The difference in the slope between log PSA and log weight for those with* $svi = 1$ *compared to those with* $svi = 0$*. To put it another way, the additional increase in average log PSA associated with a 1 unit increase in log weight for those with SVI=1 compared to those with SVI=0.*

$$\begin{aligned}\beta_3 &= E\left[log(PSA) \mid lweight, SVI = 1\right] - E\left[log(PSA) \mid lweight - 1, SVI = 1\right] \\ &\quad - \left(E\left[log(PSA) \mid lweight, SVI = 0\right] - E\left[log(PSA) \mid lweight - 1, SVI = 0\right]\right)\end{aligned}$$

3. Suppose we wish to interpret the coefficients on the original scale for PSA and weight. Write down the model in terms of the variables on their original scale, and hence, provide interpretations of the associations between PSA, weight and SVI.

   You may find it useful to repeat the plot you created in the first part, for the variables on their original scale.

   Note that the logs in this problem have base e. Please see pages 207-209 of the textbook for more information.

   We began with the model:

$$ln(psa) = \beta_0 + \beta_1 svi + \beta_2 ln(weight) + \beta_3 ln(weight) svi + \epsilon$$

   This is equivalent to:

$$\begin{aligned}psa &= exp\left(\beta_0 + \beta_1 svi + \beta_2 ln(weight) + \beta_3 ln(weight) svi\right) \\ &= exp\left(\beta_0 + \beta_1 svi\right) exp\left(ln(weight^{\beta_2}\right) exp\left(ln(weight^{\beta_3 svi})\right) \\ &= exp\left(\beta_0 + \beta_1 svi\right) weight^{\beta_2 + \beta_3 svi} \exp(epsilon)\end{aligned}$$

   Treating everything besides ϵ as a constant,

$$E[psa] = exp\left(\beta_0 + \beta_1 svi\right) weight^{\beta_2 + \beta_3 svi} E[\exp(epsilon)].$$

   Since $E[exp(\epsilon)] \neq exp[E[\epsilon]]$, just knowing that the ϵ are mean 0 is not enough to know that $E[\exp(epsilon)] = 1$.

   If the original ϵs have median 0, then the $exp(\epsilon)$ have median 1 (since $exp()$ is monotonic), and we can say that

$$Median[psa] = exp\left(\beta_0 + \beta_1 svi\right) weight^{\beta_2 + \beta_3 svi}.$$

   Therefore, some students might wish to give their interpretations below in terms of the median PSA instead of the expected PSA (and those are fine- but some assumption on the ϵ should be noted!!)

The interpretations below are in terms of expected values, but they depend on an assumption that the distribution of $\epsilon \mid X$ is constant across X. Regardless of if you interpreted in terms of medians or expected values, you should be sure that you have the correct assumptions.

- $exp(\beta_0)$ is the median psa for an individual with svi=0 and weight=1 (so that $ln(weight) = 0$).
- Under an assumption that the errors have the same distribution across X values, $exp(\beta_1)$ is the multiplicative change in (median/expected) PSA between two individuals with $weight = 1$, where one has $svi = 0$ and the other has $svi = 1$. This is because, under the assumption that the ϵ all have the same distribution regardless of X, the ϵ terms cancel in the expression below:

$$\frac{E[psa \mid svi = 1, weight = 1]}{E[psa \mid svi = 0, weight = 1]} = \frac{exp\,(\beta_0 + \beta_1)\, E[exp(\epsilon \mid svi = 1, weight = 1)]}{exp\,(\beta_0)\, E[exp(\epsilon \mid svi = 1]0, weight = 1])} = exp(\beta_1)$$

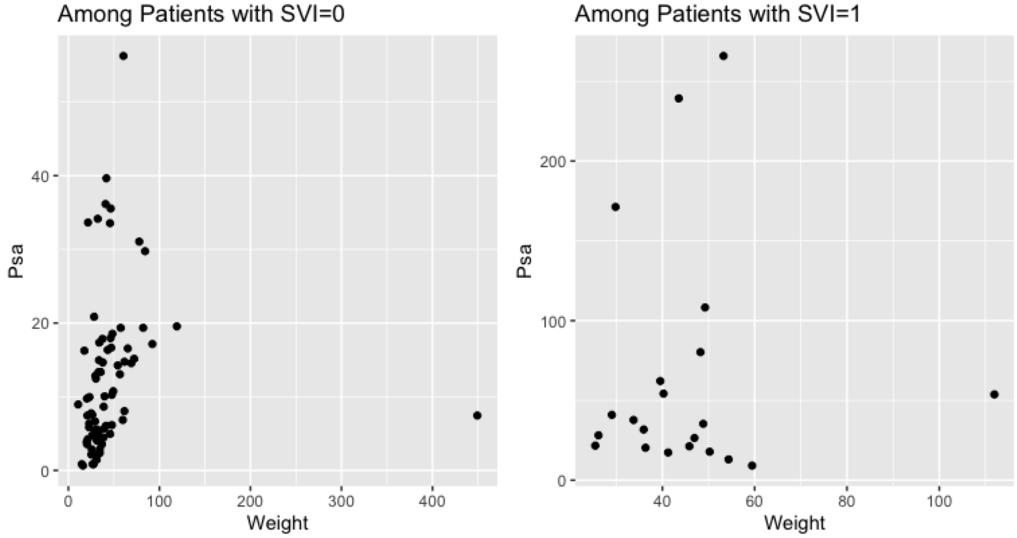
- For any positive constant c, c^{β_2} is the multiplicative increase in (median/expected) PSA associated with a c-fold change in weight for those with svi=0. When the have the same distribution:

$$\frac{E[psa \mid svi = 0, weight = cw]}{E[psa \mid svi = 0, weight = w]} = \frac{exp\,(\beta_0)\, c^{\beta_2} w^{\beta_2}}{exp\,(\beta_0)\, w^{\beta_2}} = c^{\beta_2}$$

- For a constant c, c^{β_3} gives the additional multiplicative change in (expected/median) PSA associated with a c-fold change in weight for those with $svi = 1$ (compared to the change we would have had if $svi = 0$).

$$\frac{E[psa \mid svi = 1, weight = cw]}{E[psa \mid svi = 1, weight = w]} = \frac{exp\,(\beta_0 + \beta_1)\, c^{\beta_2+\beta_3} w^{\beta_2+\beta_3}}{exp\,(\beta_0 + \beta_1)\, w^{\beta_2+\beta_3}} = c^{\beta_2} c^{\beta_3}$$

  If we then divide this whole quantity by its equivalent when $svi = 0$, we get c^{β_3}. Informally, c^{β_3} is the multiplicative change in “slope” for those with $svi = 0$ compared to those with $svi = 1$.



The assumptions part

State the assumptions that are required valid for:

1. An unbiased estimate of β_j, $j = 0, \ldots, 3$.
   We have assumed the model $Y = \boldsymbol{X}\beta + \epsilon$ and are estimating using the OLS estimator $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^TY$. Note that, E($\hat{\beta}$) $= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$E($Y|\boldsymbol{X}$) and that our estimate $\hat{\beta}$ is always unbiased for the parameter on the right. So in this case, $\hat{\beta}_1$ is unbiased for the true linear-best-fit line between $E[log(psa) \mid log(weight)]$ and and $E[log(weight)]$ among the svi=0 population. Whether or not this parameter is an interesting quantity depends on whether or not the linear model is a good fit, but that is a separate issue. If β is just defined as the population version of the least squares line, we get an unbiased estimate with no assumptions (i.e. β need not be equal to $E[Y \mid X]$.

2. An accurate estimate of the standard error of β_j, $0 = 1, \ldots, 3$.
   With no assumptions, we can say that

   $$\begin{aligned} Cov(\hat{\beta}|\boldsymbol{X}) &= Cov((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^TY \mid \boldsymbol{X}) \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Cov(Y|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{aligned}$$

   In the first part of the exercise, we then assumed that $Cov(Y \mid \boldsymbol{X}) = \sigma_\epsilon^2 I_n$ so that the expression above would simplify to $\sigma_y^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. This assumption allowed us to consistently estimate $Cov(Y \mid \boldsymbol{X})$ (which has the standard errors of the $\hat{\beta}$s on the diagonal) by plugging in $RSS/(n-p-1)$ as an estimate of σ_ϵ^2. But assuming that $Cov(Y \mid \boldsymbol{X}) = \sigma_\epsilon^2 I_n$

relies on homoskedasticity and independence of the errors ϵ. If this assumption is not met, we need a different way to estimate $Cov(Y|\boldsymbol{X})$ in order to get SE estimates for the $\hat{\beta}$s.

3. Accurate coverage probabilities for 100$(1-\alpha)$% confidence intervals of the form

   $$\widehat{\beta}_j \pm \widehat{\text{var}}(\widehat{\beta}_j)^{1/2} \times z_{1-\alpha/2},$$

   where $z_{1-\alpha/2}$ represents the $(1-\alpha/2)$ quantile of an $N(0,1)$ random variable.

   For accurate coverage from this confidence interval, we first need $\hat{\beta}$ to be an unbiased estimator and $\widehat{\text{Var}}(\hat{\beta}_j)$ to be a consistent estimator for its variance. Additionally, we need that $\hat{\beta}$ is approximately normally distributed. Approximate normality can be achieved with a reasonably large sample size (n=97 is almost certainly large enough), assuming mild regularity conditions on the distribution of X, regardless of the true distribution of the errors (so long as its variance is finite). In general, we DO NOT need the error terms to be normally distributed.

4. Accurate coverage probabilities for 100$(1-\alpha)$% confidence intervals of the form

   $$\widehat{\beta}_j \pm \widehat{\text{var}}(\widehat{\beta}_j)^{1/2} \times t_{n-4}(1-\alpha/2),$$

   where $t_{n-4}(1-\alpha/2)$ represents the $(1-\alpha/2)$ quantile of a standard Student's t random variable on $n-4$ degrees of freedom.
   Similarly, for accurate coverage from this interval we need unbiased $\hat{\beta}$, a consistent estimator of the variance, and approximate normality. Note that (iii) and (iv) are asymptotically equivalent. If the errors are in fact Normally distributed, the formula will provide exact coverage even in small samples.

5. An accurate prediction for an *observed* outcome at $\boldsymbol{x} = \boldsymbol{x}_0$.
   We must be in a situation where we can accurately estimate $E[Y \mid \boldsymbol{X}_0]$ and so the assumed mean model must be correct. Additionally, if we wanted to get correct coverage for confidence intervals for these predictions, we would need the errors to be normally distributed.

```
library(lasso2)
data(Prostate)
attach(Prostate)
y <- Prostate$lpsa
lweight <- Prostate$lweight
svi <- Prostate$svi
```

```
lmod <- lm(y~svi+lweight+svi:lweight)
summary(lmod)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5965     0.7127  -0.837 0.404727
svi           4.1826     2.4149   1.732 0.086591 .
lweight       0.7541     0.1946   3.876 0.000198 ***
svi:lweight  -0.7197     0.6425  -1.120 0.265521

Residual standard error: 0.8969 on 93 degrees of freedom
```

1 Appendix: Code for plots

```
library(ggplot2)
library(patchwork)
library(dplyr)
p1 <- ggplot(data=Prostate %>% filter(svi==0), aes(x=lweight, y=lpsa))+geom_point()+
   geom_smooth(method=lm, se=FALSE)+ggtitle("Among Patients with SVI=0")+
   xlab("log weight")+ylab("log psa")
p2 <- ggplot(data=Prostate %>% filter(svi==1), aes(x=lweight, y=lpsa))+geom_point()+
   geom_smooth(method=lm, se=FALSE)+ggtitle("Among Patients with SVI=1")+
   xlab("log weight")+ylab("log psa")
p1+p2

p1 <- ggplot(data=Prostate %>% filter(svi==0), aes(x=exp(lweight), y=exp(lpsa)))+
   geom_point()+ggtitle("Among Patients with SVI=0")+
   xlab("Weight")+ylab("Psa")
p2 <- ggplot(data=Prostate %>% filter(svi==1), aes(x=exp(lweight), y=exp(lpsa)))+
   geom_point()+ggtitle("Among Patients with SVI=1")+
   xlab("Weight")+ylab("Psa")
p1+p2
```