Problem Set 4 Solution

Math 430, Winter 2017

Problem 1

The data in `heart.txt` are from a study of the effect of a new drug on a particular heart function. These drugs have been developed for another purpose, but one concern is whether they have a side effect on heart function. Drugs A and B are two forms of the drug, C is a placebo (i.e. a control, expected to have no effect on heart function). Thirty subjects were randomly assigned to a drug. The intent was to have 10 subjects per drug, but a mistake was made and drug B was given instead of drug C to one of the subjects. PRE is the heart function measured before the drug was administered. POST is the heart function 2 hours after the drug was administered.

```
heart <- read.table("https://raw.githubusercontent.com/math430-lu/data/master/heart.txt", header = TRUE
```

(a)

Consider only the post drug data. Is there evidence of an effect of the drugs on heart function?

```
heart_mod1 <- lm(post ~ drug, data = heart)
summary(heart_mod1)
```

```
##
## Call:
## lm(formula = post ~ drug, data = heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5000 -3.3523  0.7955  2.9432 14.5000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   75.500      1.572  48.021   <2e-16 ***
## drugB          2.409      2.172   1.109    0.277
## drugC          2.167      2.284   0.948    0.351
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.972 on 27 degrees of freedom
## Multiple R-squared:  0.05058,    Adjusted R-squared:  -0.01975
## F-statistic: 0.7192 on 2 and 27 DF,  p-value: 0.4962
```

There is no evidence of an effect of the drugs on heart function, as seen by the F-statistic of 0.72 with associated p-value of 0.4962.

(b)

Estimate the post-drug means for each treatment, the mean difference between drug A and the placebo (C), and the s.e. of that difference.

To see how to obtain these estimates, consider the below conditional expectations:

$E(post|drug = A) = \beta_0$

$E(post|drug = B) = \beta_0 + \beta_1$

$E(post|drug = C) = \beta_0 + \beta_2$

- The estimated post-drug mean for treatment A: $\widehat{\beta}_0 = 75.5$
- The estimated post-drug mean for treatment B: $\widehat{\beta}_0 + \widehat{\beta}_1 = 75.5 + 2.409 = 77.9$
- The estimated post-drug mean for treatment C: $\widehat{\beta}_0 + \widehat{\beta}_2 = 75.5 + 2.167 = 77.7$
- The mean difference between drug A and the placebo (C) is given by β_2, so we estimate it to be $\widehat{\beta}_2 = 2.167$, and the s.e. is $se(\widehat{\beta}_2) = 2.284$.

(c)

Consider a model that uses PRE drug data as a predictor. Assume a linear regression with the same slope for all drugs. Is there evidence of an effect of the drugs on heart function? Report your test statistic and p-value.

To answer this question, we can use a partial F-test to compare the two models:

```r
heart_mod2 <- lm(post ~ drug + pre, data = heart)
heart_mod3 <- lm(post ~ pre, data = heart)
anova(heart_mod3, heart_mod2)
```

```
## Analysis of Variance Table
##
## Model 1: post ~ pre
## Model 2: post ~ drug + pre
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     28 460.83
## 2     26 360.40  2    100.43 3.6226 0.04094 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes, there evidence of an effect of the drugs on heart function. F = 3.62 with p-value = 0.041.

(d)

Estimate the post-drug means for each treatment, when pre-treatment function is set to the overall mean.

```r
pre_mean <- mean(heart$pre)
predict(heart_mod2, newdata = data.frame(drug = "A", pre = pre_mean))
```

```
##        1
## 74.58284
```

```r
predict(heart_mod2, newdata = data.frame(drug = "B", pre = pre_mean))
```

```
##        1
## 77.37092
```

```r
predict(heart_mod2, newdata = data.frame(drug = "C", pre = pre_mean))
```

```
##       1
## 79.3435
```

(e)

Estimate the mean difference between drug A and the placebo (C), for subjects with the same PRE heart function. Also report the s.e. of the adjusted difference.

Note that

$E(post|drug = A, pre = x) = \beta_0 + \beta_3 x$

$E(post|drug = C, pre = x) = \beta_0 + \beta_2 + \beta_3 c$

```
broom::tidy(heart_mod2)
```

```
##          term   estimate  std.error statistic      p.value
## 1 (Intercept) 12.1324794 13.5160806  0.897633 3.776186e-01
## 2       drugB  2.7880833  1.6287354  1.711809 9.883714e-02
## 3       drugC  4.7606587  1.7972567  2.648847 1.355121e-02
## 4         pre  0.8337832  0.1771672  4.706194 7.299583e-05
```

The difference between A and C is estimated to be 4.8 (This is for C - A) with s.e. = 1.8.

(f)

One of the major assumptions of the above model is that the slope—i.e. the relationship between PRE and the response—is the same for all three drugs. Is this assumption reasonable here?

A parital F-test comparing the parallel line and unrelated line models can be used:

```
heart_mod4 <- lm(post ~ pre * drug, data = heart)
anova(heart_mod2, heart_mod4)
```

```
## Analysis of Variance Table
##
## Model 1: post ~ drug + pre
## Model 2: post ~ pre * drug
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 360.40
## 2     24 344.27  2    16.125 0.5621 0.5774
```

There is no evidence of different slopes. The p-value for the F-test is 0.58.

Alternative: You could have looked at the t-test for the interaction pre*drug, which also had a p-value of 0.58.

(g)

One of your office mates finds some of your results rather curious. Please explain (briefly) why the differences in b) and e) are not the same number. Also, explain why the se's are different.

The estimated differences are not the same because drugs A and C have different mean pretreatment values and the slope for pre is not zero. For A, the mean pre = 76, but for C, the mean pre = 72.9. The standard errors are different because the covariate accounts for a moderate amount of the variability in post-treatment responses. The s.d. of the errors drops from 4.97 without the covariate to 3.72 with the covariate.

Problem 2

The data file, `salary.csv` contains salary and other characteristics of all faculty in a small Midwestern college collected in the early 1980s for presentation in legal proceedings for which discrimination against women in salary was at issue. All persons in the data set hold tenured or tenure-track positions; temporary faculty are not included. The variables included are outlined in the table below.

Variable	Description
`degree`	`PhD` or `MS`
`rank`	`Asst`, `Assoc`, or `Prof`
`sex`	`Male` or `Female`
`Year`	number of years in the current rank

Variable	Description
ysdeg	years since earning their highest degree
salary	academic year salary in dollars

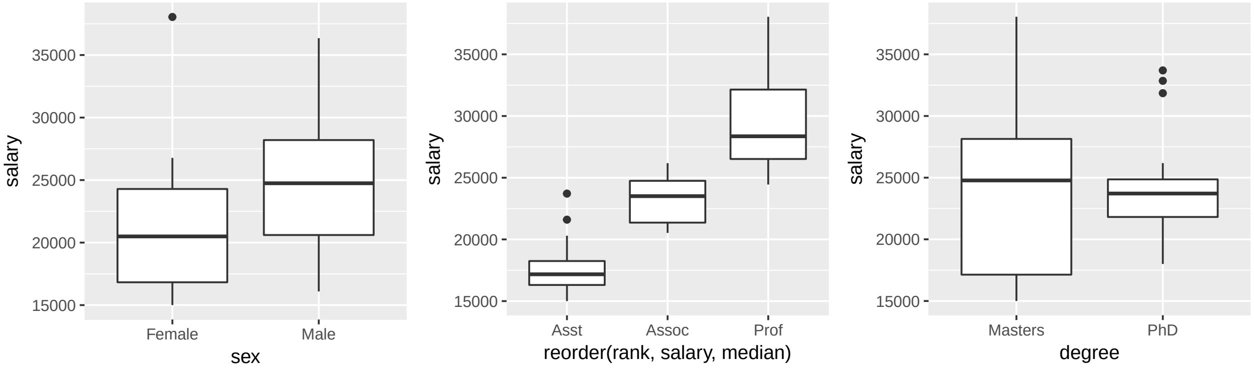
```
salary <- read.csv("https://raw.githubusercontent.com/math430-lu/data/master/salary.csv")
library(ggplot2)
```

(a)

Create appropriate graphical summaries exploring the data and discuss your findings.

Let's start with the factors. These can be displayed in boxplots:

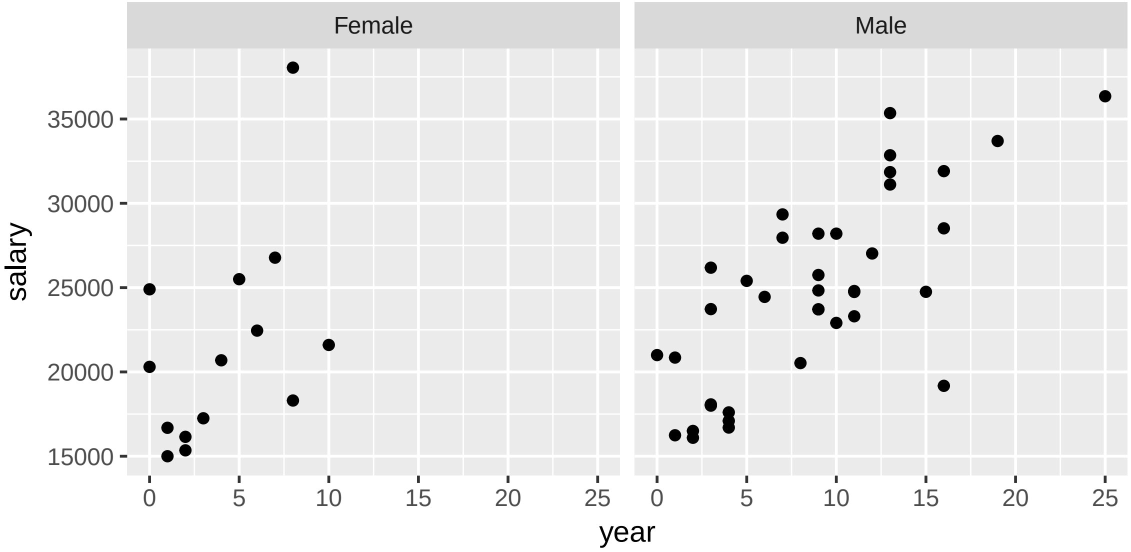
```
ggplot(data = salary, aes(x = sex, y = salary)) +
  geom_boxplot()

ggplot(data = salary, aes(x = rank, y = salary)) +
  geom_boxplot()

ggplot(data = salary, aes(x = degree, y = salary)) +
  geom_boxplot()
```



Female salaries appear to be generally lower than male salaries, salary increases with rank. Faculty with a Masters degree have much more variable salaries. The boxplots don't have anything to say about interactions.

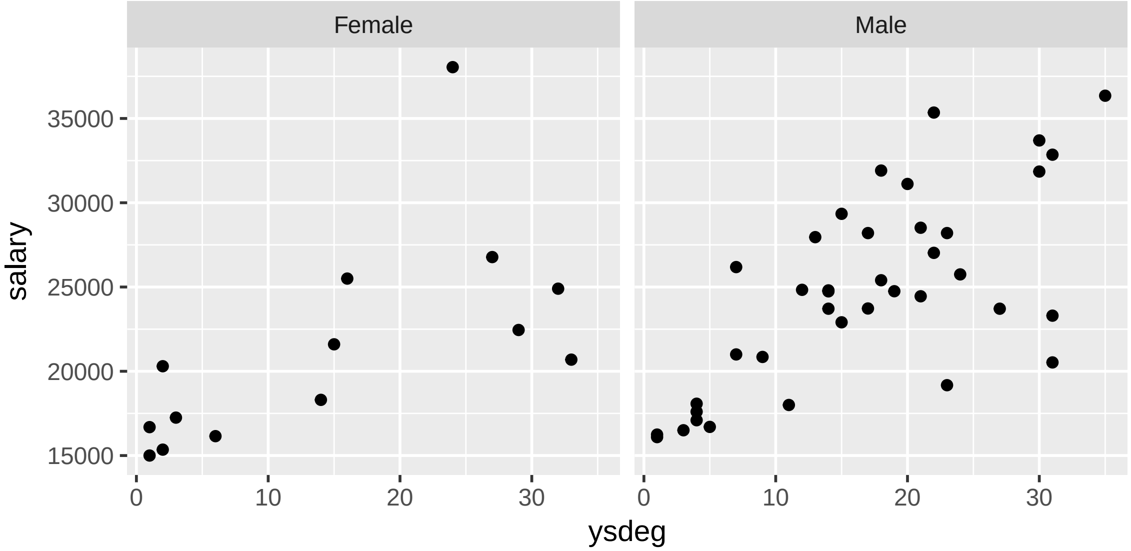
Turn next to the two continuous variables:

```
ggplot(data = salary, aes(x = year, y = salary)) +
  geom_point() +
  facet_wrap(~ sex)
```



Females generally have fewer years in rank, and while for males salary clearly increases with year, this is not so clear for females.

```r
ggplot(data = salary, aes(x = ysdeg, y = salary)) +
  geom_point() +
  facet_wrap(~ sex)
```



Interestingly, the females are more variable on `ysdeg` than on `year`. For this variable it does appear that salary increases with `ysdeg` for both sexes.

(b)

Test the hypothesis that the mean salary for men and women is the same. Be careful to specify a logical alternative hypothesis.

This can be tested using regression software by fitting an intercept and a dummy variable for sex.

```r
salary_mod1 <- lm(salary ~ sex, data = salary)
summary(salary_mod1)
```

```
##
## Call:
## lm(formula = salary ~ sex, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8602.8 -4296.6  -100.8  3513.1 16687.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    21357       1545  13.820   <2e-16 ***
## sexMale         3340       1808   1.847   0.0706 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5782 on 50 degrees of freedom
## Multiple R-squared:  0.0639, Adjusted R-squared:  0.04518
## F-statistic: 3.413 on 1 and 50 DF,  p-value: 0.0706
```

We are testing $H_0 : \beta_1 = 0$ vs. $H_0 : \beta_1 > 0$ (remember that discrimination against women is hypothesize). Based on the p-value of $0.0706/2 = 0.0353$ (because we have a one-sided test), there is evidence that women are paid less on average than men.

(c)

Assuming no interactions between `sex` and the other predictors, obtain a 95% confidence interval for the mean difference in salary between males and females.

```
salary_mod2 <- lm(salary ~ ., data = salary)
confint(salary_mod2, parm = "sexMale")
```

```
##             2.5 %   97.5 %
## sexMale -3030.565 697.8183
```

Adjusting for the other predictors, the `sex` effect is for higher salaries for females, although there is no evidence of a difference as the confidence interval contains zero.

(d)

Finkelstein (1980), in a discussion of the use of regression in discrimination cases wrote, "[a] variable reflect a position or status bestowed by the employer, in which case if there is discrimination in the award of position or status, the variable may be 'tainted.'" Thus, for example, if discrimination is at work in promotion of faculty to higher ranks, using rank to adjust salaries before comparing the sexes may not be acceptable to the courts.

Exclude the `rank` variable, refit the model from part (c), and summarize your findings.

```
salary_mod3 <- lm(salary ~ . - rank, data = salary)
summary(salary_mod3)
```

```
##
## Call:
## lm(formula = salary ~ . - rank, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8146.9 -2186.9  -491.5  2279.1 11186.6
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15897.03    1259.87  12.618  < 2e-16 ***
## degreePhD   -3299.35    1302.52  -2.533 0.014704 *
## sexMale      1286.54    1313.09   0.980 0.332209
## year          351.97     142.48   2.470 0.017185 *
## ysdeg         339.40      80.62   4.210 0.000114 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3744 on 47 degrees of freedom
## Multiple R-squared:  0.6312, Adjusted R-squared:  0.5998
## F-statistic: 20.11 on 4 and 47 DF,  p-value: 1.048e-09
```

If we ignore rank, then the coefficient for Sex is again positive, indicating an advantage for males, but the p-value is .33 (or .165 for a one-sided test), indicating that the difference is not significant.

One could argue that other variables in this data set are tainted as well, so using data like these to resolve issues of discrimination will never satisfy everyone.