HW3

Jordan Garrett

10/31/2019

1)

a.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

b.

$$\hat{\beta} = \left(\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\frac{1}{n(\sum_{i=1}^n x_i - \overline{x})^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -n\overline{x} \\ -n\overline{x} & n \end{bmatrix} \begin{bmatrix} n\overline{y} \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

$$\frac{1}{(\sum_{i=1}^n x_i - \overline{x})^2} \begin{bmatrix} \overline{y} \sum_{i=1}^n x_i^2 - \overline{x} \sum_{i=1}^n x_i y_i \\ -\overline{x} n \overline{y} + \sum_{i=1}^n x_i y_i \end{bmatrix}$$

$$\frac{1}{(\sum_{i=1}^n x_i - \overline{x})^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 - y\overline{x}^2 + y\overline{x}^2 - \overline{x} \sum_{i=1}^n x_i y_i \\ (\sum_{i=1}^n x_i y_i - \overline{xy}) \end{bmatrix}$$

$$\frac{1}{(\sum_{i=1}^n x_i - \overline{x})^2} \begin{bmatrix} \overline{y}(\sum_{i=1}^n x_i^2 - \overline{x}^2) - \overline{x}(\sum_{i=1}^n x_i y_i - \overline{yx}) \\ \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}) \end{bmatrix}$$

$$\begin{bmatrix} \overline{y} - \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{(\sum_{i=1}^n x_i - \overline{x})^2} \overline{x} \\ \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{(\sum_{i=1}^n x_i - \overline{x})^2} \end{bmatrix}$$

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \overline{y} - \hat{\beta}_2 \overline{x} \\ \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{(\sum_{i=1}^n x_i - \overline{x})^2} \end{bmatrix}$$

c.

$$\frac{\sigma^2}{n(\sum_{i=1}^n x_i - \overline{x})^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -n\overline{x} \\ -n\overline{x} & n \end{bmatrix}$$

If the off-diagonal elements of the covaraince matrix are 0, then the β are uncorrelated. Since $n \neq 0$, then the only way for the off-diagonal elements to be uncorrelated is if $\overline{x} = 0$. I would reformulate the model by subtracting $\overline{x}$ from it.

2)

a.

Model:

$$\begin{bmatrix} y_{1,1} \\ y_{1,2} \\ y_{2,1} \\ y_{2,2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{1,1} \\ \epsilon_{1,2} \\ \epsilon_{2,1} \\ \epsilon_{2,2} \end{bmatrix}$$

Least Square Estimates:

$$\begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} = \begin{bmatrix} \frac{y_{1,1}+y_{1,2}}{2} \\ \frac{y_{2,1}+y_{2,2}}{2} \end{bmatrix}$$

Hypothesis Test:

$$H_0 : A\hat{\theta} = 0$$

$$\begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \frac{y_{1,1}+y_{1,2}}{2} \\ \frac{y_{2,1}+y_{2,2}}{2} \end{bmatrix} = 0$$

$$F = \frac{(\begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \frac{y_{1,1}+y_{1,2}}{2} \\ \frac{y_{2,1}+y_{2,2}}{2} \end{bmatrix})^T (\begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \frac{y_{1,1}+y_{1,2}}{2} \\ \frac{y_{2,1}+y_{2,2}}{2} \end{bmatrix})}{RSS/2}$$

$$F = \frac{(y_{1,1} + y_{1,2} - y_{2,1} - y_{2,2})^2/4}{RSS/2}$$

$$F = \frac{(y_{1,1} + y_{1,2} - y_{2,1} - y_{2,2})^2/2}{|| \begin{bmatrix} y_{1,1} - \frac{y_{1,1}+y_{1,2}}{2} \\ y_{1,2} - \frac{y_{1,1}+y_{1,2}}{2} \\ y_{2,1} - \frac{y_{2,1}+y_{2,2}}{2} \\ y_{2,2} - \frac{y_{2,1}+y_{2,2}}{2} \end{bmatrix} ||^2}$$

$$F = \frac{(y_{1,1} + y_{1,2} - y_{2,1} - y_{2,2})^2/2}{((y_{1,1} - y_{1,2})^2 + (y_{2,1} - y_{2,2})^2)/2}$$

$$F = \frac{(y_{1,1} + y_{1,2} - y_{2,1} - y_{2,2})^2}{(y_{1,1} - y_{1,2})^2 + (y_{2,1} - y_{2,2})^2}$$

b.

Model:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}$$

Least Squares Estimate:

$$\begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} = \begin{bmatrix} \frac{2y_1+y_3-y_2}{3} \\ \frac{2y_2+y_3-y_1}{3} \end{bmatrix}$$

Hypothesis Test:

$$H_0 : A\hat{\theta} = 0$$

$$F = \frac{\left(\begin{bmatrix} 1 - 1 \end{bmatrix} \begin{bmatrix} \frac{2y_1+y_3-y_2}{3} \\ \frac{2y_2+y_3-y_1}{3} \end{bmatrix} \right)^T \left[\begin{bmatrix} 1 - 1 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right]^{-1} \left(\begin{bmatrix} 1 - 1 \end{bmatrix} \begin{bmatrix} \frac{2y_1+y_3-y_2}{3} \\ \frac{2y_2+y_3-y_1}{3} \end{bmatrix} \right)}{RSS/(3-2)}$$

$$F = \frac{(y_1 - y_2)^2/2}{RSS}$$

$$F = \frac{(y_1 - y_2)^2/2}{|| \begin{bmatrix} y_1 - (2y_1 - y_2 + y_3)/3 \\ y_2 - (-y_1 + 2y_2 + y_3)/3 \\ y_3 - (y_1 + y_2 + 2y_3)/3 \end{bmatrix} ||^2}$$

$$F = \frac{(y_1 - y_2)^2/2}{(y_1 + y_2 - y_3)^2/3}$$

3)

a.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \\ y_{n+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \\ \epsilon_{n+1} \end{bmatrix}$$

b.

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \overline{y}_n \\ y_{n+1} \end{bmatrix}$$

c.

Hypothesis Testing:

$$H_0 : A\hat{\beta} = 0$$

$$F = \frac{(\overline{y_n} - y_{n+1})^2/(\frac{1}{n} + 1)}{RSS/(n-1)}$$

$$F = \frac{(\overline{y_n} - y_{n+1})^2/(\frac{1}{n} + 1)}{|| \begin{bmatrix} y_1 - \overline{y}_n \\ y_2 - \overline{y}_n \\ \vdots \\ y_n - \overline{y}_n \\ y_{n+1} - y_{n+1} \end{bmatrix} ||^2/(n-1)}$$

$$F = \frac{(\overline{y_n} - y_{n+1})^2/(\frac{1}{n} + 1)}{S_n^2/(n-1)}$$

$$F = \frac{n(\overline{y_n} - y_{n+1})^2(n-1)}{S_n^2(n+1)}$$

4)

a.

```r
teengamb$sex <- factor(teengamb$sex, labels = c("M","F"))
describe(teengamb)
```

```
##         vars  n  mean    sd median trimmed   mad  min max range  skew
## sex*       1 47  1.40  0.50   1.00    1.38  0.00  1.0   2   1.0  0.38
## status     2 47 45.23 17.26  43.00   45.28 22.24 18.0  75  57.0  0.10
## income     3 47  4.64  3.55   3.25    4.14  2.45  0.6  15  14.4  1.33
## verbal     4 47  6.66  1.86   7.00    6.79  1.48  1.0  10   9.0 -0.79
## gamble     5 47 19.30 31.52   6.00   12.90  8.75  0.0 156 156.0  2.35
##         kurtosis   se
## sex*       -1.90 0.07
## status     -1.31 2.52
## income      1.10 0.52
## verbal      0.69 0.27
## gamble      5.97 4.60
```

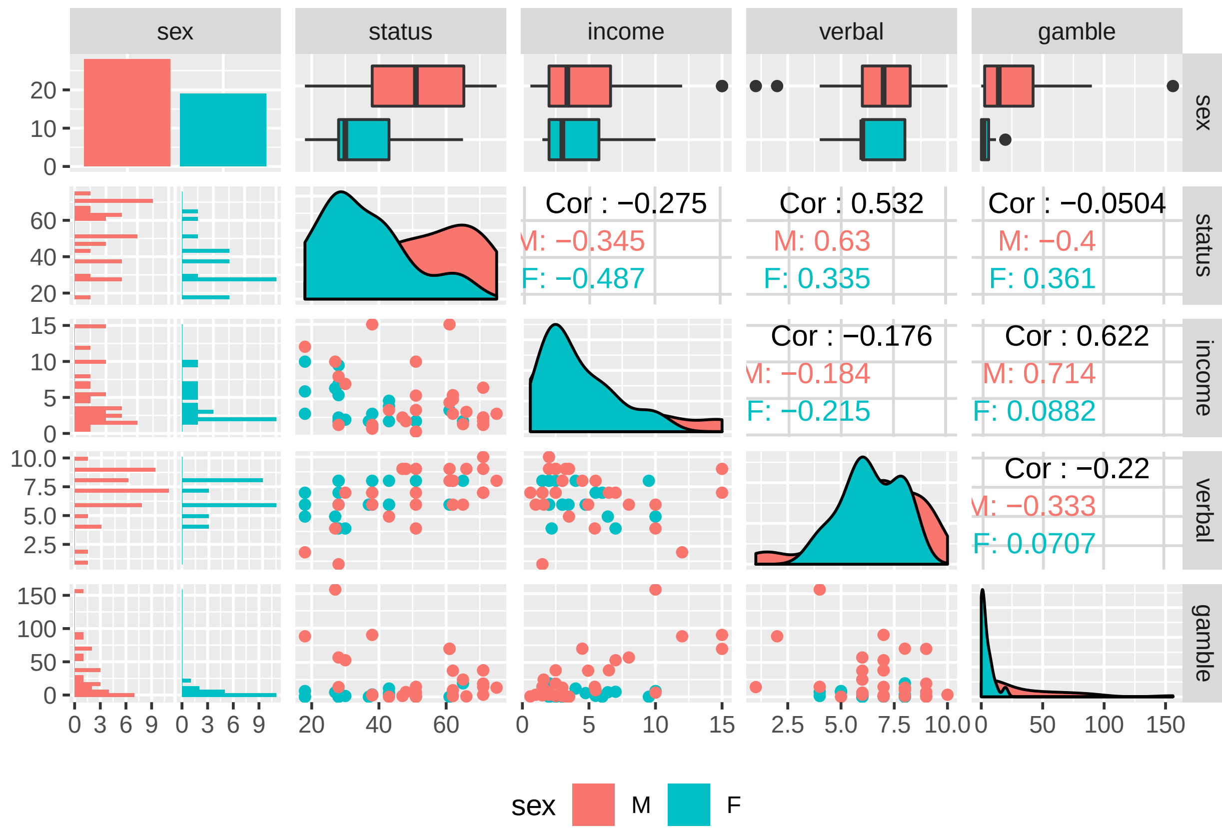
```r
describeBy(teengamb[,c(2:5)],teengamb$sex)
```

```
##
##  Descriptive statistics by group
## group: M
##         vars  n  mean    sd median trimmed   mad  min max range  skew
## status     1 28 52.00 16.43  51.00   52.71 19.27 18.0  75  57.0 -0.37
## income     2 28  4.98  4.09   3.38    4.49  2.78  0.6  15  14.4  1.16
## verbal     3 28  6.82  2.14   7.00    7.04  1.48  1.0  10   9.0 -0.95
## gamble     4 28 29.77 37.32  14.25   24.48 19.94  0.0 156 156.0  1.60
##         kurtosis   se
## status     -1.14 3.11
## income      0.27 0.77
## verbal      0.51 0.41
## gamble      2.37 7.05
## ---------------------------------------------------------
## group: F
##         vars  n  mean    sd median trimmed   mad  min  max range  skew
## status     1 19 35.26 13.43   30.0   34.53 11.86 18.0 65.0  47.0  0.67
## income     2 19  4.15  2.60    3.0    3.96  1.48  1.5 10.0   8.5  0.94
## verbal     3 19  6.42  1.35    6.0    6.47  1.48  4.0  8.0   4.0 -0.23
## gamble     4 19  3.87  5.15    1.7    3.17  2.52  0.0 19.6  19.6  1.60
##         kurtosis   se
## status     -0.44 3.08
## income     -0.37 0.60
## verbal     -1.16 0.31
## gamble      2.08 1.18
```

```r
ggpairs(teengamb, aes(color=sex), legend=c(1,1)) + theme(legend.position = "bottom")
```

```
## `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`.
```



- There are a couple of things about this data stands out to me:
  - There are a lot more men than women.
  - None of the continuous variables are normally distributed, rather all are skewed.
  - Status is strongly correlated with verbal, while income is strongly correlated with gambling. These strong correlations can become a problem if each of these variables are used as predictors.

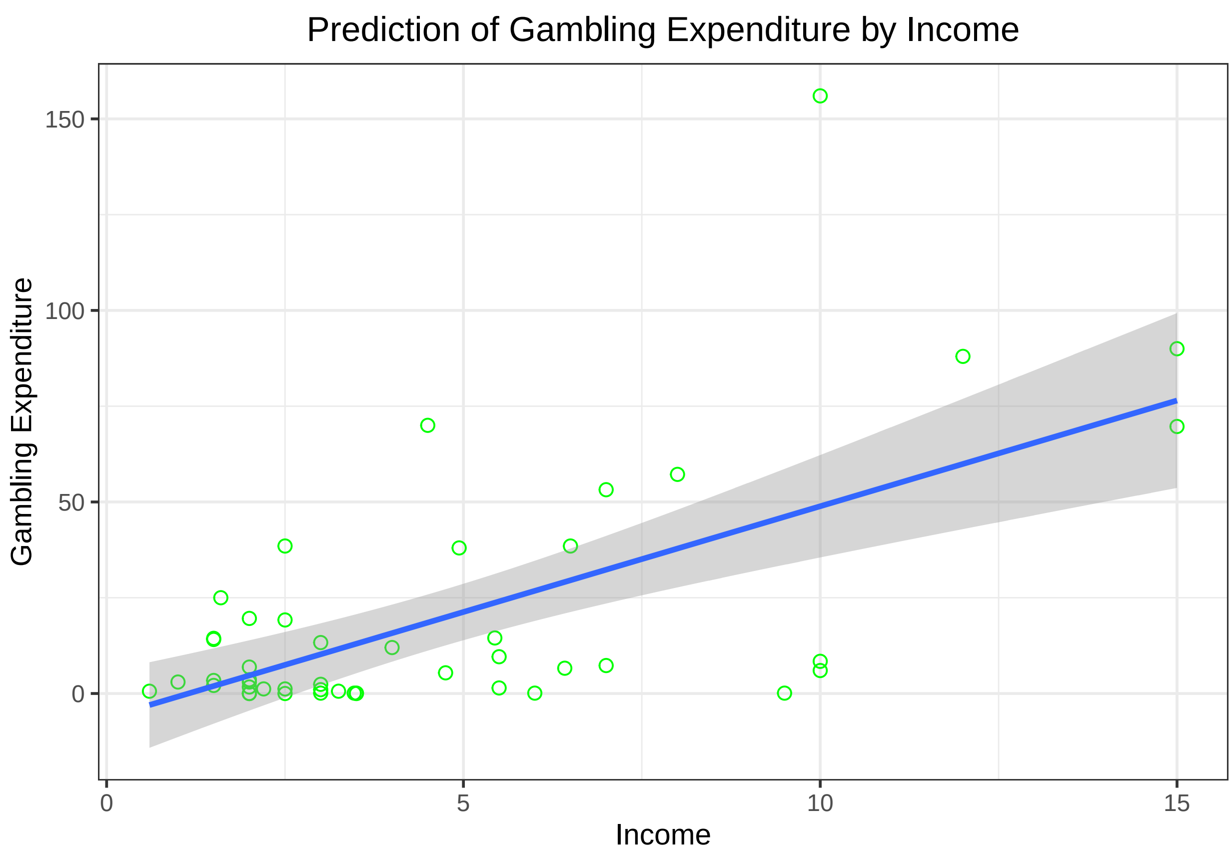
b.

```r
summary(fit.1 <- lm(gamble~income, data=teengamb))
```

```
##
## Call:
## lm(formula = gamble ~ income, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.020 -11.874  -3.757  11.934 107.120
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.325      6.030  -1.049      0.3
## income         5.520      1.036   5.330 3.05e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.95 on 45 degrees of freedom
## Multiple R-squared:  0.387,  Adjusted R-squared:  0.3734
## F-statistic: 28.41 on 1 and 45 DF,  p-value: 3.045e-06
```

```
ggplot(teengamb, aes(x=income,y=gamble)) +
  geom_point(shape=21, color="green", size=2) +
  geom_smooth(method='lm') +
  labs(title="Prediction of Gambling Expenditure by Income",
       x="Income", y="Gambling Expenditure")+
  theme_bw()+theme(plot.title = element_text(hjust = 0.5))
```



Results of the fit indicate a significant positive relationship between income and gambling expenditure ($t(45) = 5.330$, $p < 0.001$). At an income of 0, gambling expenditure is -6.325, although this term was not significant. With every one unit increase in income (β_2), gambling expenditure increased on average by 5.520 units. The magnitude of the effect was large according to Cohen's guidelines; income explained 38.7% of the variance in gambling expenditure (R^2=0.387).

c.

```
X <- cbind(rep(1,nrow(teengamb)),teengamb$income)
y <- teengamb$gamble
Beta.hat <- solve(t(X)%*%X)%*%t(X)%*%y
Beta.hat
```

```
##           [,1]
## [1,] -6.324559
## [2,]  5.520485
```

LS estimates are the same as derived using the lm() function.

d.

Income explained 38.7% of the variance in gambling expenditure (R^2=0.387).

e.

```
largest_Resid.case <- which(abs(fit.1$residuals) == max(abs(fit.1$residuals)))
largest_Resid.case
```

```
## 24
## 24
```

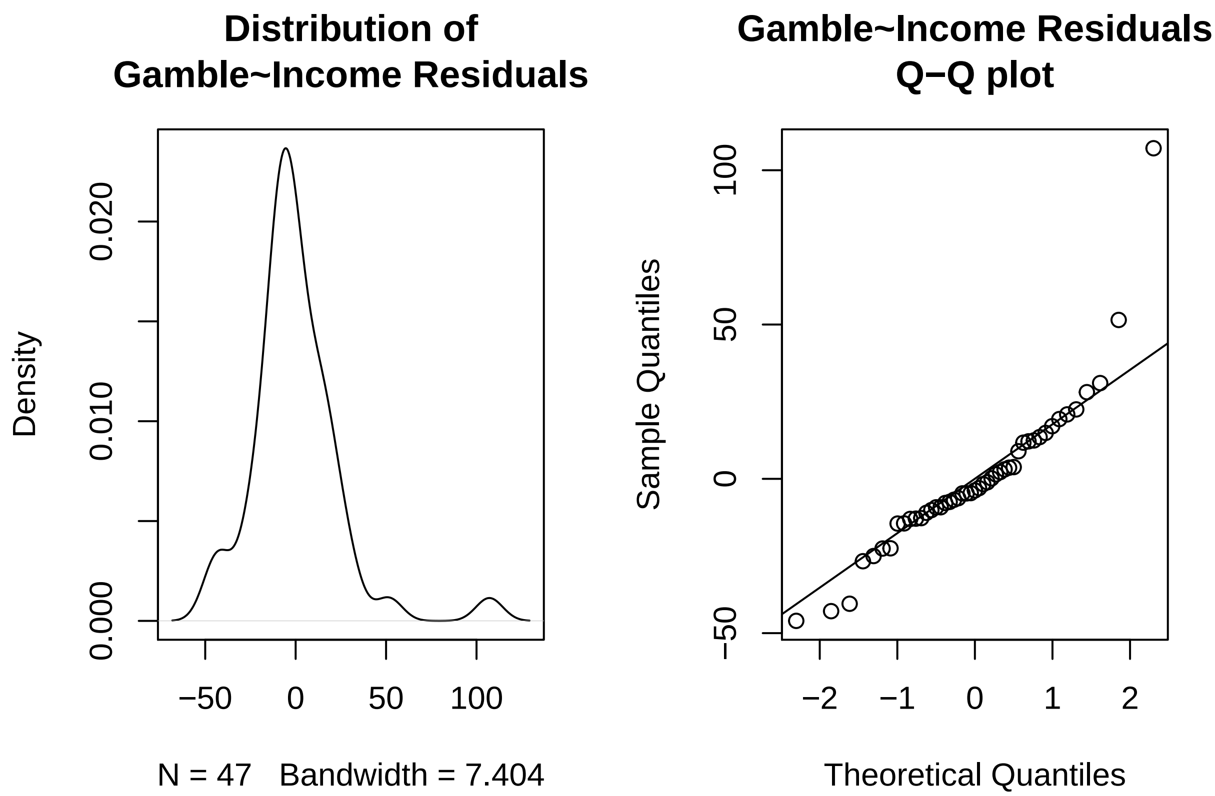
Case number 24 has the largest absolute residual (107.1197).

f.

```
summary(fit.1$residuals)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -46.020 -11.874  -3.757   0.000  11.934 107.120
```

```
par(mfrow=c(1,2))
plot(density(fit.1$residuals), main="Distribution of\nGamble~Income Residuals")
qqnorm(fit.1$residuals, main="Gamble~Income Residuals\nQ-Q plot")
qqline(fit.1$residuals)
```



The linear model assumes that the residuals have a mean of 0, and are normally distributed. The mean of our residuals is in fact 0, while our median is slightly smaller. There is a clear outlier in our residual data, which may be influencing the mean and making it incongruent to the median. If we were to remove this outlier, then both the mean and median may = 0.

g.

```
sqrt(summary(fit.1)$r.squared)
```

```
## [1] 0.6220769
```

The multiple correlation coefficient between gambling expenditure and income is R=0.6220769. This is the square root of the amount of variance in gambling expenditure explained by income.

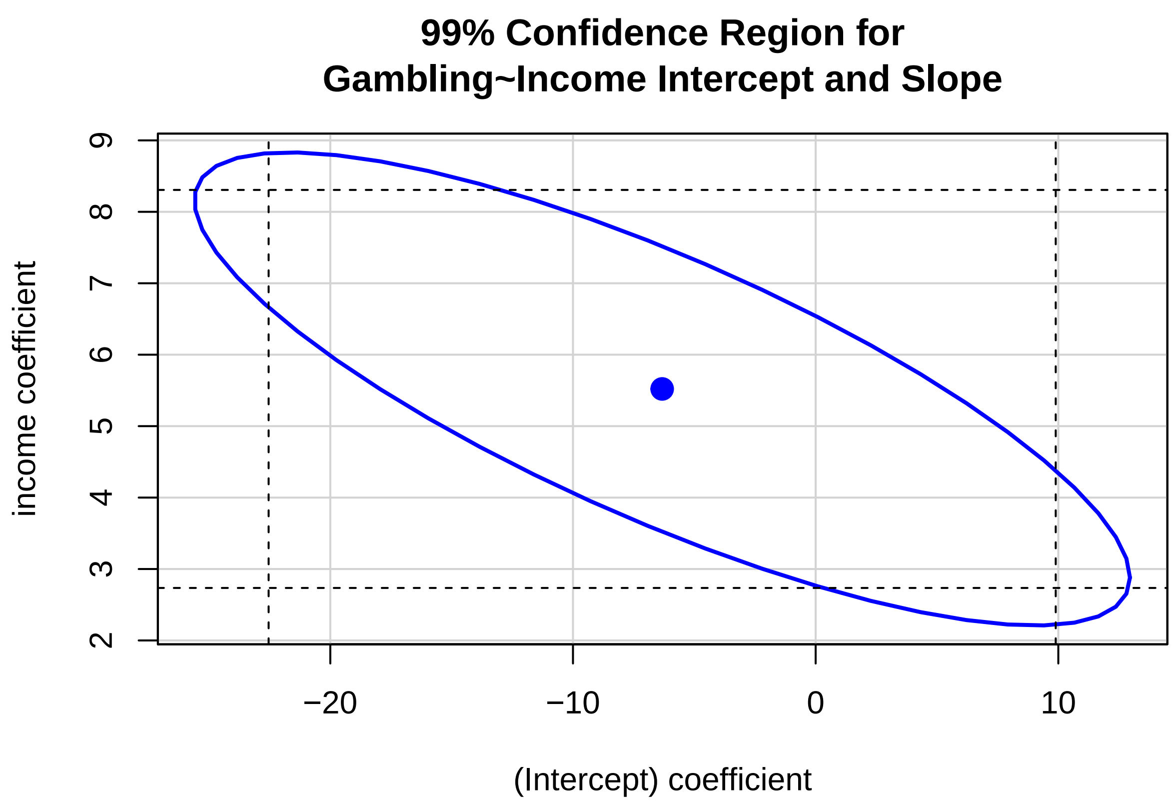
h.

```
confint(fit.1,level=.99)
```

```
##                  0.5 %   99.5 %
## (Intercept) -22.542419 9.893300
## income        2.734687 8.306283
```

i.

```
confidenceEllipse(fit.1, level=.99,
                  main = "99% Confidence Region for\nGambling~Income Intercept and Slope")
abline(v=confint(fit.1, level=.99)[1,], lty=2)
abline(h=confint(fit.1, level=.99)[2,], lty=2)
```

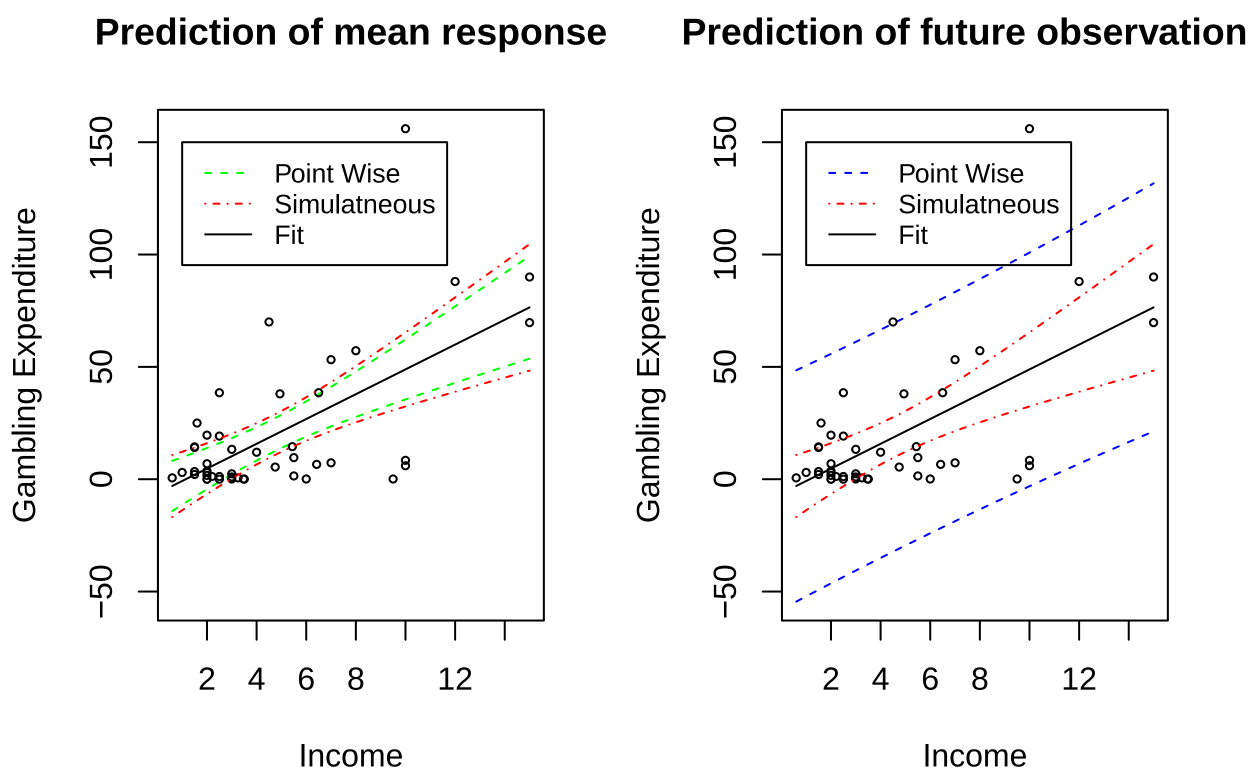


j.

```
grid <- seq(min(teengamb$income),max(teengamb$income),len=100)
p1 <- predict(fit.1, newdata=data.frame(income=grid), se=T, level=.95,
             interval="confidence")
p2 <- predict(fit.1, newdata=data.frame(income=grid), se=T, level=.95,
             interval="prediction")
par(mfrow=c(1,2))
matplot(grid,p1$fit,lty=c(1,2,2),col=c("black","green","green"),type="l",
  xlab="Income",ylab="Gambling Expenditure",
  ylim=range(p1$fit,p2$fit,teengamb$gamble))
points(teengamb$income,teengamb$gamble,cex=.5)
title("Prediction of mean response")
lines(grid,
        p1$fit[,1]-sqrt(2*qf(.95,2,length(grid)-2))*p1$se.fit,
```

```
         lty=4, col="red")
lines(grid,
         p1$fit[,1]+sqrt(2*qf(.95,2,length(grid)-2))*p1$se.fit,
         lty=4, col="red")
legend(1,150,legend=c("Point Wise","Simulatneous","Fit"), col=c("green","red","black"),
        lty = c(2,4,1),cex=0.8)


matplot(grid,p2$fit,lty=c(1,2,2),col=c("black","blue","blue"),type="l",
  xlab="Income",ylab="Gambling Expenditure",
  ylim=range(p1$fit,p2$fit,teengamb$gamble))
points(teengamb$income,teengamb$gamble,cex=.5)
title("Prediction of future observations")
lines(grid,
         p2$fit[,1]-sqrt(2*qf(.95,2,length(grid)-2))*p2$se.fit,
         lty=4, col="red")
lines(grid,
         p2$fit[,1]+sqrt(2*qf(.95,2,length(grid)-2))*p2$se.fit,
         lty=4, col="red")
legend(1,150,legend=c("Point Wise","Simulatneous","Fit"), col=c("blue","red","black"),
        lty = c(2,4,1),cex=0.8)
```

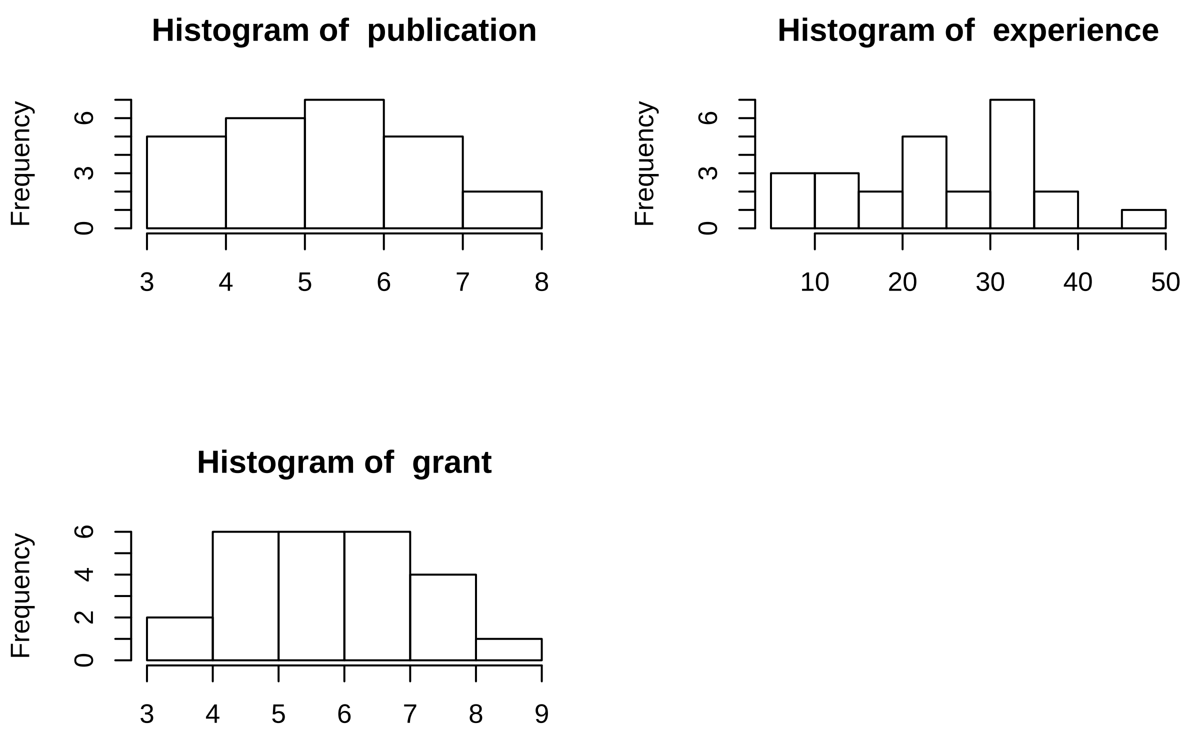


5)

```r
math.salaryData <- read.csv("/Users/owner/Downloads/salary_data.csv",sep=" ", header=F)
colnames(math.salaryData) <- c("publication","experience","grant","salary")
```

a.

```r
par(mfrow=c(2,2))
for (i in 1:3) {
  hist(math.salaryData[,i],xlab="",
       main=paste("Histogram of ",names(math.salaryData)[i]))
}
```

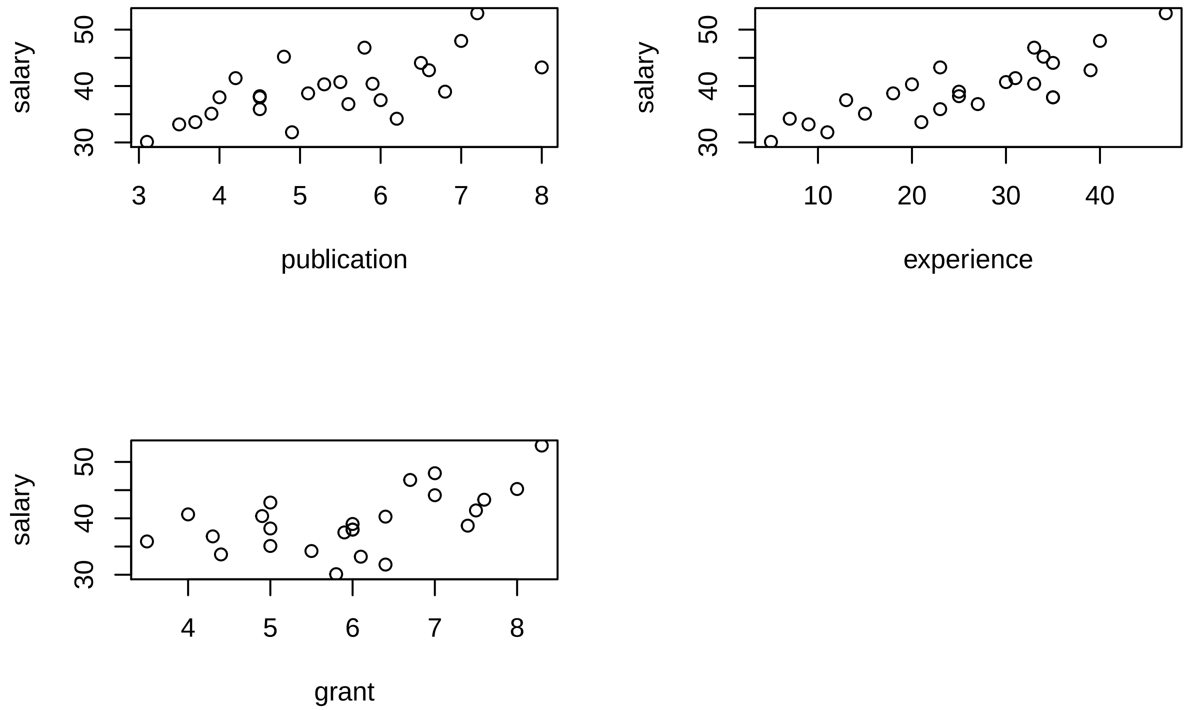


The variables publication and grant are somewhat normally distributed, while experience slighty has a skew.

b.

```r
par(mfrow=c(2,2))
for (i in 1:3){
  plot(math.salaryData[,i],math.salaryData$salary,
```

```r
        xlab=names(math.salaryData[i]), ylab="salary")
}
```



Publication quality and years of experience qualitatively have the strongest positive relationship with annual salary, while grant support has a slightly less positive relationship.

c.

```r
summary(fit.2 <- lm(salary~publication+experience+grant, data=math.salaryData))
```

```
##
## Call:
## lm(formula = salary ~ publication + experience + grant, data = math.salaryData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3261 -1.0274 -0.1519  1.2361  3.5426
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.40780    2.13249   8.163 5.95e-08 ***
## publication  1.26031    0.34324   3.672 0.001420 **
## experience   0.30179    0.03837   7.865 1.08e-07 ***
## grant        1.28073    0.31980   4.005 0.000642 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.878 on 21 degrees of freedom
## Multiple R-squared:  0.8929, Adjusted R-squared:  0.8776
## F-statistic: 58.34 on 3 and 21 DF,  p-value: 2.344e-10
```

d.

The model fits salary data significantly well. The combination of the three predictors accounts for 89.3% of the variance in annual salary (R^2=0.893; F(3,21)=58.34, $p < 0.001$)

e.

Slope estimates for publication quality ($\beta_1 = 1.26$, $p < 0.01$), years of experience ($\beta_2 = 0.30$, $p < 0.001$) and grant support ($\beta_3 = 1.2$, $p < 0.001$) were all significant. Total amount of variance explained (see part d) was also significant.

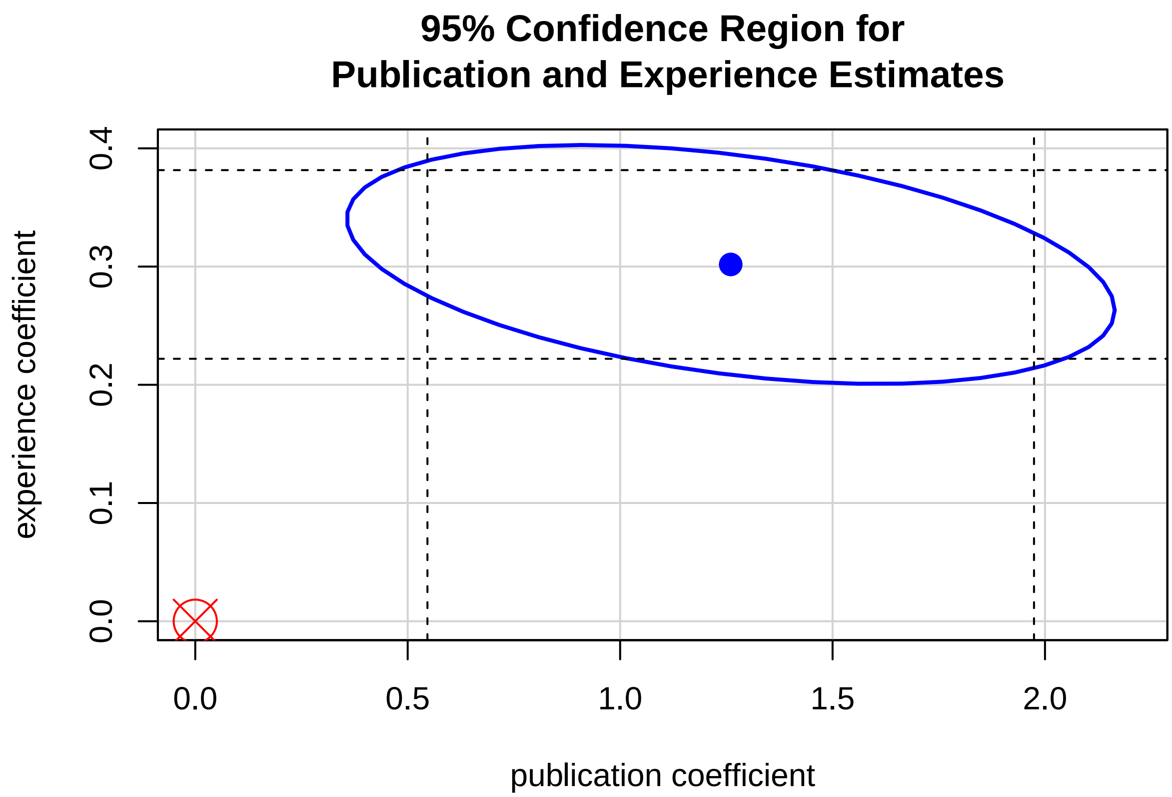
f.

```r
confint(fit.2, level=.90)[3,]
```

```
##       5 %      95 %
## 0.2357596 0.3678147
```

For x_2 $p < 0.05$, but for x_1 this confidence interval does not inform us about its p-value.

```r
confidenceEllipse(fit.2,c(2,3) , level=.95,
                  main = "95% Confidence Region for\n Publication and Experience Estimates",
                  ylim=c(0,0.4), xlim=c(0,2.2))
abline(v=confint(fit.2, level=.95)[2,], lty=2)
abline(h=confint(fit.2, level=.95)[3,], lty=2)
points(0,0, col='red', cex=3, pch=13)
```



The origin of the confidence represents the null hypothesis of $\beta - \hat{\beta} = 0$. The confidence region takes into account correlations between each β since they come from the same dataset, and if the origin is not within the bounds of the confidence region then we can reject the aformentioned null hypothesis. Observing the plot above, it is clear that the origin is not near the confidence region, thus indicating that we can reject the null hypothesis $\beta - \hat{\beta} = 0$.

h.

```
new_data <- data.frame("publication"=c(4,5,6,7),
                       "experience" = c(10,20,30,50),
                       "grant" = c(4,5,6,7))
```

```
scheffe(fit.2,new_data)
```

```
##        fit      lwr      upr
## 1 30.58981 27.86048 33.31915
## 2 36.14872 34.48384 37.81360
## 3 41.70763 40.22146 43.19380
## 4 50.28441 47.11573 53.45308
```

i.

```r
p4 <- predict(fit.2, newdata = data.frame("publication"=7,
                                           "experience"=10,
                                           "grant"=7.9),
              se=T, interval="prediction")
p4$fit
```

```
##        fit      lwr      upr
## 1 39.36558 34.76018 43.97098
```

They are not grossly underpaid because the 95% confidence interval includes their current salary, although a pay bump of $4k would be nice.