Lecture 5 Exercises

Isabel Fulcher

8/13/2018

Install packages

```
library(tidyverse) #ggplot2, dplyr, etc.
library(reshape2) #need this for melt()
library(MASS) #contains dataset
```

Exercise I

Load the birthwt data. This data contains 189 observations, 9 predictors, and an outcome, birthweight, available both as a continuous measure and a binary indicator for low birth weight.

```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622
```

1. Plot a scatterplot of birthweight (bwt) and mother's weight (lwt).

```
p <- ggplot(birthwt,aes(x=lwt,y=bwt)) + geom_point() +
  labs(x="Mother's weight (pounds)",y="Birthweight (grams)")
```

2. Use OLS to fit the regression of birthweight on mother's weight.

```
# Fit the regression and view results
fit <- lm(bwt ~ lwt, data=birthwt)
#This returns a summary of the results
summary(fit)
```

3. Extract the following: estimated coefficients, standard errors, variance-covariance matrix, and confidence intervals.

```
# Estimated coefficients
coefficients(fit)

# Standard errors for the above
summary(fit)$coeff[,2]

# Variance-covariance matrix
vcov(fit)

# Confidence intervals
# use this funtion (do not recommend using for glm())
confint(fit)
```

4. Plot the regression line and interpret the intercept and slope

```r
p <- ggplot(birthwt,aes(x=lwt,y=bwt)) +
  geom_point(color="grey") + #plots gray points
  stat_smooth(method="lm",col="red",se=FALSE) #plot regression line
```

5. Does the interpretation of the intercept make sense? How might we change this?

```r
birthwt %>% mutate(lwt_star = lwt - mean(lwt)) -> birthwt

fit.new <- lm(bwt ~ lwt_star,data=birthwt)
summary(fit.new)
```

6. Now, we want to fit a model that includes race, mother's age, and smoking status in the model. Race takes on value 1 for white, 2 for black, and 3 for other. Mother's age is continuous. Smoking status is binary. Write out the regression function we may be interested in.
7. Use OLS to calculate the coefficient estimates in this model.

```r
fit2 <- lm(bwt ~ as.factor(race) + age + smoke, data=birthwt)
summary(fit2)
```

```
##
## Call:
## lm(formula = bwt ~ as.factor(race) + age + smoke, data = birthwt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2322.6  -447.3    28.4   502.2  1612.3
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3281.673    260.664  12.590  < 2e-16 ***
## as.factor(race)2 -444.069    156.194  -2.843 0.004973 **
## as.factor(race)3 -447.858    119.017  -3.763 0.000226 ***
## age                 2.134      9.771   0.218 0.827326
## smoke            -426.093    109.988  -3.874 0.000149 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 690 on 184 degrees of freedom
## Multiple R-squared:  0.1236, Adjusted R-squared:  0.1046
## F-statistic:  6.49 on 4 and 184 DF,  p-value: 6.592e-05
```

8. Interpret all the coefficient estimates.

- The estimated mean birthweight among infants born to mothers who are non-smokers are zero years old (weird!), and white is 3281.7 grams.
- The estimated mean birthweight among infants with black mothers is 444.07 grams lower than the mean birthweight among infants with white mothers, holding all other variables constant
- The estimated mean birthweight among infants with mothers in the "other" race category is 444.86 grams lower than the mean birthweight among infants with white mothers, holding all other variables constant
- The estimated change in mean birthweight corresponding to a one year change in mother's age is 2.134 grams, holding all other variables constant
- The estimated mean birthweight among infants with mothers that smoke is 426.09 grams lower than the mean birthweight among infants with mothers that do not smoke, holding all other variables constant.

9. Print the results in Rmarkdown using kable().

```
table <- data.frame(summary(fit2)$coef)
row.names(table) <- c("Intercept","White","Black","Mother's age", "Smoker")

knitr::kable(table,digits=3,align=rep('c', 2),
      col.names = c("estimate","standard error","test statistic","p-value"))
```

	estimate	standard error	test statistic	p-value
Intercept	3281.673	260.664	12.590	0.000
White	-444.069	156.194	-2.843	0.005
Black	-447.858	119.017	-3.763	0.000
Mother's age	2.134	9.771	0.218	0.827
Smoker	-426.093	109.988	-3.874	0.000

Group Exercises

From the course website, load the North Carolina infant mortality dataset. This contains information on all 225,152 births in North Carolina from 2003-2004.

```
load("infants.dat")
```

Group 1

The goal of this exercise is to emulate "sampling" from this North Carolina birth population and see how variability in our estimates will change with sample size.

1. You are interested in how maternal age affects birthweight. Write the form of the linear regression model.

$$E[Y|X] = \beta_0 + \beta_1 X$$

2. Take a sample of size $n = 100$ from this population, fit the linear regression from 1 and extract the coefficient estimate for gestational age, i.e. $\hat{\beta}_1$.

```
dat <- infants[sample(nrow(infants),100),]
fit <- lm(weight ~ weeks,data=dat)
beta <- coefficients(fit)[2]
```

3. Repeat part 2 $b = 500$ times and plot the estimated coefficients in a histogram.

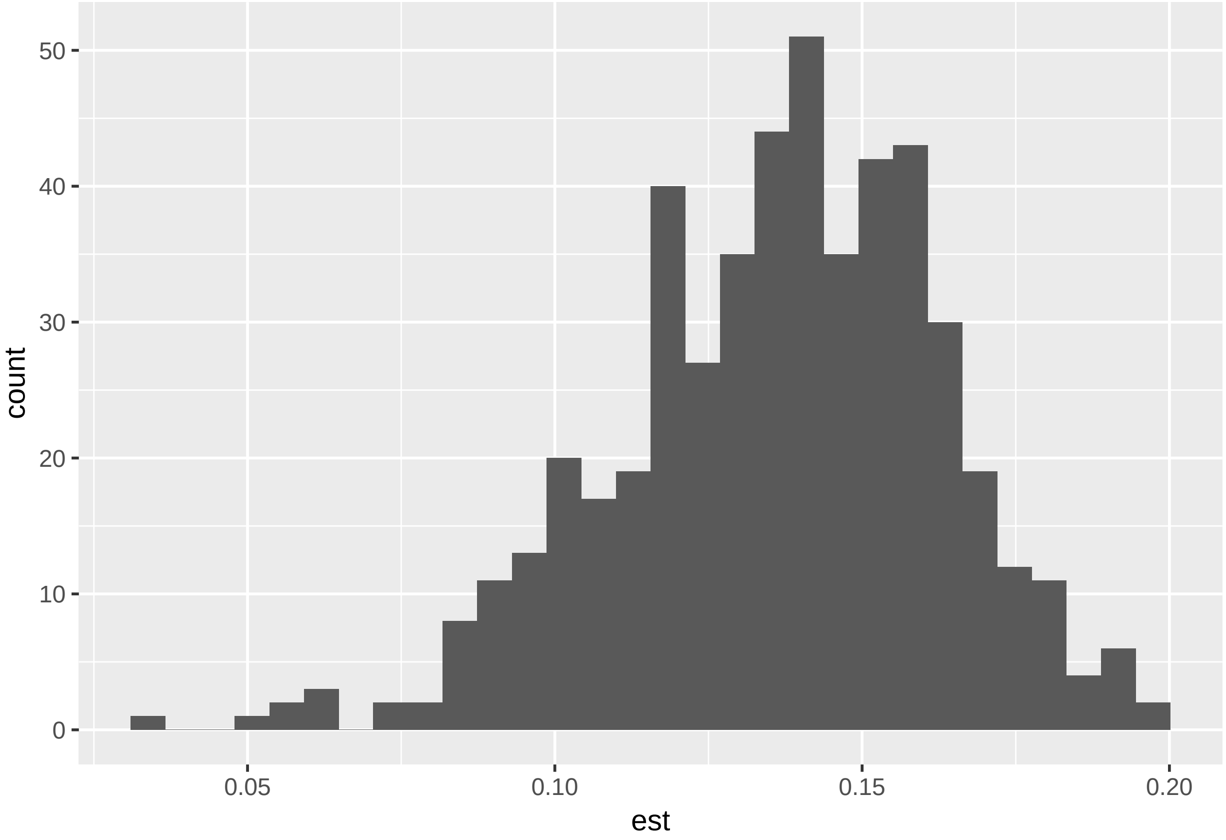
```
# Create a function to apply over
fit_reg <- function(n,b){
  dat <- infants[sample(nrow(infants),n),]
  fit <- lm(weight ~ weeks,data=dat)
  beta <- coefficients(fit)[2]
  return(beta)
}

# Apply over function
out <- sapply(1:500,fit_reg,n=100)

# Plot histogram
out.df <- data.frame(out) #make dataframe for ggplot
```

```
colnames(out.df) <- "est" #rename columns
ggplot(out.df,aes(x=est)) + geom_histogram()
```

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



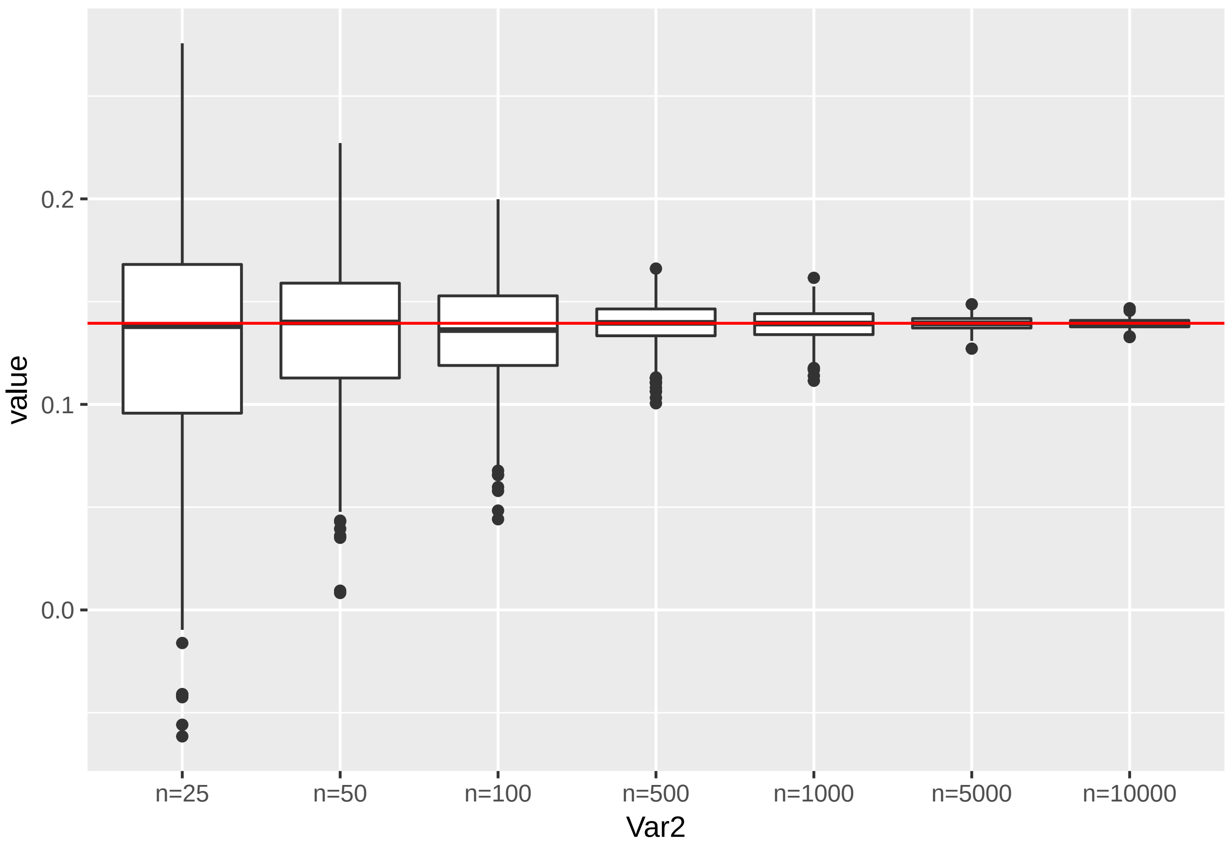
4. For the following sample sizes, $N={25, 50, 100, 500, 1000, 5000, 10000}, repeat questions 2-3. Save in a dataframe so you can plot your results.

```
B <- 500
N = c(25,50,100,500,1000,5000,10000)
results <- sapply(N, function(n) {sapply(1:B,fit_reg,n=n)} )
colnames(results) <- paste0("n=",N)
results.df <- melt(results)
```

5. Find a creative way to plot your results, and include some reference to the population β_1. Interpret these results.

```
#population beta
beta.truth <- coefficients(lm(weight ~ weeks, data=infants))[2]

ggplot(results.df,aes(x=Var2,y=value)) + geom_boxplot() +
  geom_hline(yintercept=beta.truth,col="red")
```

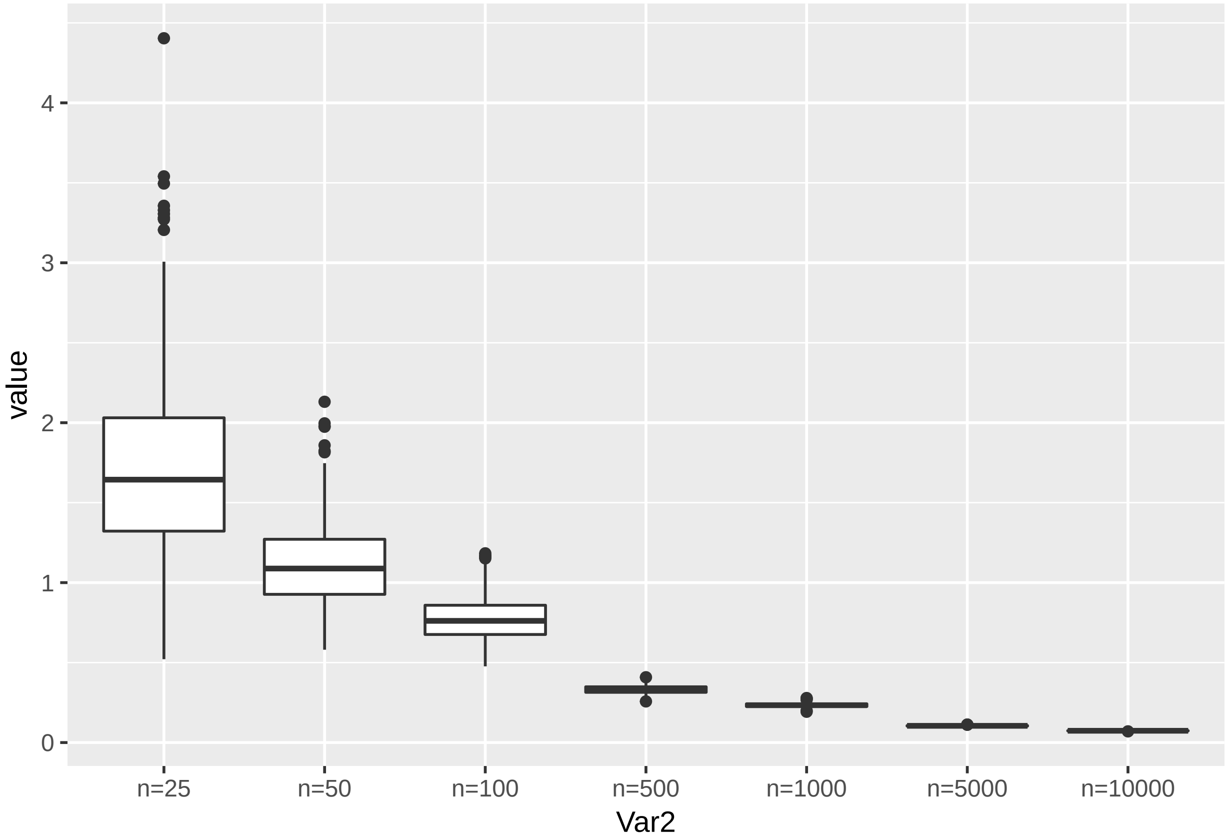


6. If you had instead extracted the standard error estimate for maternal age, what would you expect to happen? Confirm your intuition by repeating this procedure again for the standard error of the beta coefficient estimate at various sample sizes.

```
# create a function
fit_sd <- function(n,b){
  dat <- infants[sample(nrow(infants),n),]
  fit <- lm(weight ~ weeks,data=dat)
  se <- summary(fit)$coefficients[1,2]
  return(se)
}

B <- 500
N = c(25,50,100,500,1000,5000,10000)
results <- sapply(N, function(n) {sapply(1:B,fit_sd,n=n)} )
colnames(results) <- paste0("n=",N)
results.df <- melt(results)

ggplot(results.df,aes(x=Var2,y=value)) + geom_boxplot()
```



Group 2

The purpose of this exercise is to practice working with interpreting regression output with interaction terms in R.

1. Take a random sample of size 10,000 from the NC dataset to work with for this problem. Make sure everyone in your group uses the same seed, so that you draw the same sample.

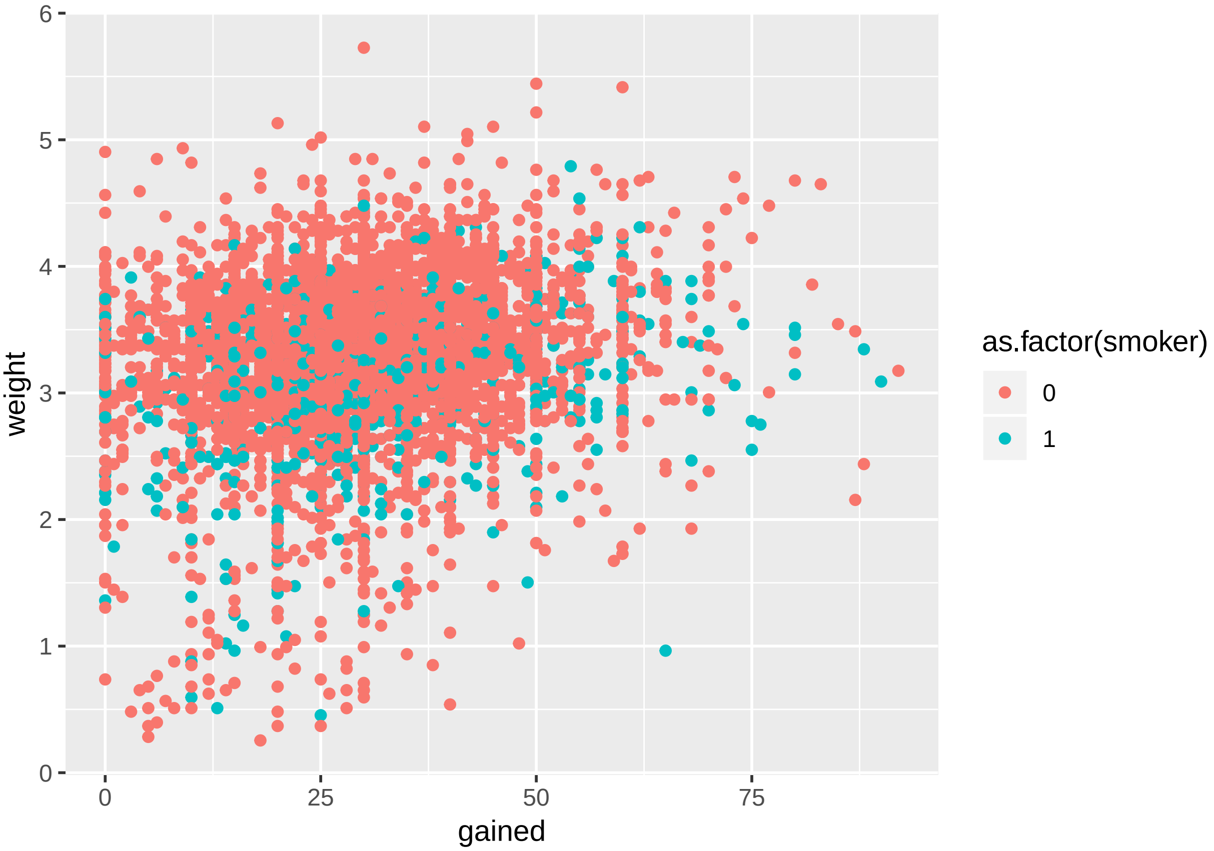
```r
set.seed(42)
dat <- infants[sample(nrow(infants),5000),]
```

2. For this problem, you will be working with the following model where Y is birth weight, X_1 is weight gain during pregnancy and X_2 is smoking. What does β_3 represent? Why might this be of interest?

$$E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

3. Create a scatter plot of maternal weight gain and birth weight. Color observations according to smoking status.

```r
ggplot(dat,aes(x=gained,y=weight,color=as.factor(smoker))) + geom_point()
```



4. Use the expression $\hat{\beta}$ given in the slides to find the estimates of the coefficients. Note: for this question, you will need to create the "design" matrix, $\mathbf{X}$.

```
dat %>% dplyr::select(gained,smoker) %>% mutate(beta0 = 1, int = gained*smoker) %>% as.matrix() -> X
dat %>% dplyr::select(weight) %>% as.matrix() -> Y

beta.hat <- solve(crossprod(X))%*%t(X)%*%Y
beta.hat
```

```
##                 weight
## gained  0.0082429763
## smoker -0.2581443107
## beta0   3.0547583271
## int     0.0007077899
```

5. Fit this regression using the lm() function. How does this compare with the results from part 4?

```
fit <- lm(weight ~ gained + smoker + I(gained*smoker),data=dat)
summary(fit)
```

```
##
## Call:
## lm(formula = weight ~ gained + smoker + I(gained * smoker), data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.94813 -0.29448  0.04552  0.37781  2.42495
##
```

```
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.0547583  0.0225702 135.345  < 2e-16 ***
## gained               0.0082430  0.0006864  12.008  < 2e-16 ***
## smoker              -0.2581443  0.0561358  -4.599 4.36e-06 ***
## I(gained * smoker)   0.0007078  0.0016182   0.437    0.662
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6088 on 4996 degrees of freedom
## Multiple R-squared:  0.04964,    Adjusted R-squared:  0.04907
## F-statistic: 86.99 on 3 and 4996 DF,  p-value: < 2.2e-16
```

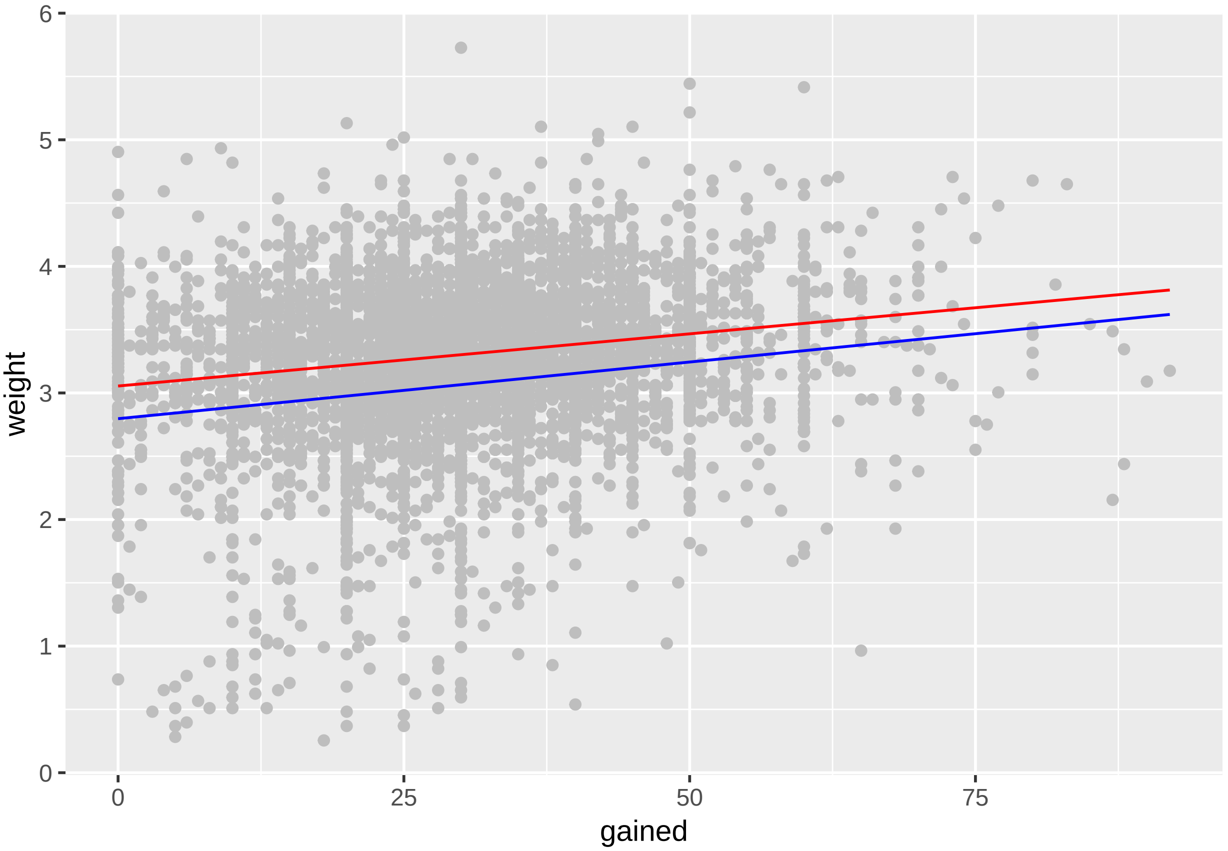
6. Interpret the coefficients for weight gain and smoker. Be as precise as possible.

- The estimated mean birthweight corresponding to a one-pound change in weight gain is 0.008 kg among non-smoking mothers.
- The estimated mean birthweight is 0.258 kg lower than among non-smokers among women who did not gain any weight.

7. Plot the regression line for smokers and non-smokers on plot 2. Hint: use stat_function() in gpplot and define your own function.

```r
beta_est_smoke <- coefficients(fit)
beta_est_nosmoke <- coefficients(fit)[c(1,2)]

ggplot(dat,aes(x=gained,y=weight)) + geom_point(color="grey") +
  stat_function(fun = function(x) {beta_est_nosmoke[1] + beta_est_nosmoke[2]*x},color="red") +
  stat_function(fun = function(x) {beta_est_smoke[1] + beta_est_smoke[2]*x +
      beta_est_smoke[3] + beta_est_smoke[4]*x},color="blue")
```



8. Do you see large differences in the slopes of these lines? Which p-value in the regression output formally tests this? Does this align with your expectations?

There is almost no difference in the slope of these lines. The p-value corresponding to the hypothesis test of $H_0 : \beta_3 = 0$ is 0.662, meaning there is a 0.662 chance that we would observe this estimated value given that the null is true.