Week 10 Live Session - Selected Answers

w203 Instructional Team

Announcements

Simple linear regression

Suppose we have data, represented by $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Q1.1: Write a simple regression model for the i^th case.

$Y_i = \beta_0 + \beta_1 X_i + u_i$, where u_i is the statistical error for the i^{th} case. The model includes this equation and a number of further assumptions that constrain the error.

Q1.2: The statistical errors u_i cannot be observed. Why?

One answer is that because we don't know the betas, meaning we don't know what the true regression line is, we also can't compute the errors.

Another answer is that the errors symbolize all of the other factors, aside from X, that can influence the outcome. We can never measure all of these.

Q1.3: What assumptions do we need to make?

There isn't just one set of assumptions for linear regression. We can use a small set of assumptions, but only get limited statistical guarantees, or if we can meet stronger assumptions, we can do more (compute standard errors, run hypothesis tests, etc).

This week, we started by presenting four assumptions:

1. Linear population Model: $Y = \beta_0 + \beta_1 X + u$
2. Random Sampling. Each (X_i,Y_i) is drawn independently from the population model.
3. $var(X) \neq 0$
4. $E(u|X) = 0$

Q1.4: Do we want the residuals to be small in magnitude? Why or why not?

Yes, residuals near zero mean that our regression line is close to the data. This corresponds to our notions of goodness of fit.

Q1.5: To define a regression line, is it sufficient to require $\Sigma\ \hat{u_i} = 0$.

No, any line through the means of the variables - the point $(\bar{X}\ \bar{Y}\ )$ satisfies $\Sigma \hat{u_i} = 0$. For example, we could draw a perfectly flat line through the mean of Y and it would satisfy this condition.

Properties of residuals

Q2.1: What are the implications of the following properties?

(1) $\sum \hat{u_i} = 0$. This really tells us that the regression line passes through the "middle of the data," the point defined by the mean of each variable $(\bar{X}\ \bar{Y}\ )$.

(2) $\Sigma\ X_i\ \hat{u_i} = 0$

The residuals are uncorrelated with the independent variables.

Q2.2: How many different lines through the X-Y plane would fulfill these two conditions?

Just one! These are what we call sample moment conditions. As Paul demonstrated in the async (watch "Method of Moments"), if you begin with these two conditions and solve the equations, you arrive at exactly the same line as you get by minimizing least squares residuals using calculus.

It helps to think about the two conditions separately. The first one forces the line you get to pass through ($\bar{X}$ $\bar{Y}$), but the line is still free to rotate around this point. As you rotate the line, however, the residuals can become more or less correlated with X. The second condition locks us down to just a single line.

Q2.2: Using the above conditions, compute $cov(\hat{Y}_i, \hat{u}_i)$.

$cov(\hat{Y}_i, \hat{u}_i) = cov(\hat{\beta}_0 + \hat{\beta}_1 X_i, \hat{u}_i) = cov(\hat{\beta}_0, \hat{u}_i) + \hat{\beta}_1 cov(X_i, \hat{u}_i) = 0 + 0 = 0$

Remember this fact: the outcome variable is also uncorrelated with the residuals.

Regression in R

When a linear pattern is evident from a scatter plot, the relationship between the two variables is often modeled with a straight line. This line is expressed in a linear model between the response (or dependent) variable and the predictor (or independent) variable.

The following functions are useful for running a linear regression in R.

- Fitting a model: model <- lm(y ~ x)
- Coefficients: model$coef or coef(model)
- Fitted values: model$fitted or fitted(model)
- Residuals: model$resid or resid(model)

Install and load the BSDA pacakge using the commands install.packages("BSDA") and library(BSDA),respectively.

We are interested in using the GPA data frame, which we can attach using the command: attach(Gpa), to investigate the impact of high school GPA on college GPA.

Before we can find the least square regression line, we need to determine the explanatory and response varibles. Define 2 new variables in R, x and Y, and assign the explanatory and response variables from the dataset, respectively, and conduct a cursory analysis of the data set.

```
library(BSDA)
```

```
## Loading required package: e1071
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'BSDA'
```

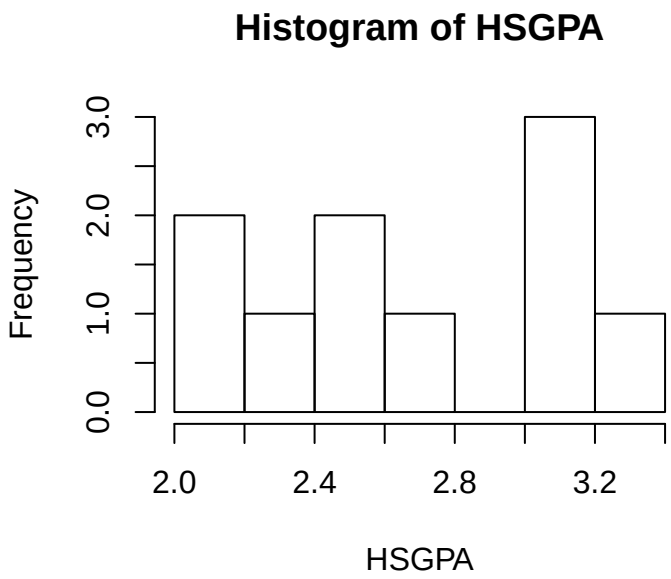
```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```
attach(Gpa)
summary(Gpa)
```

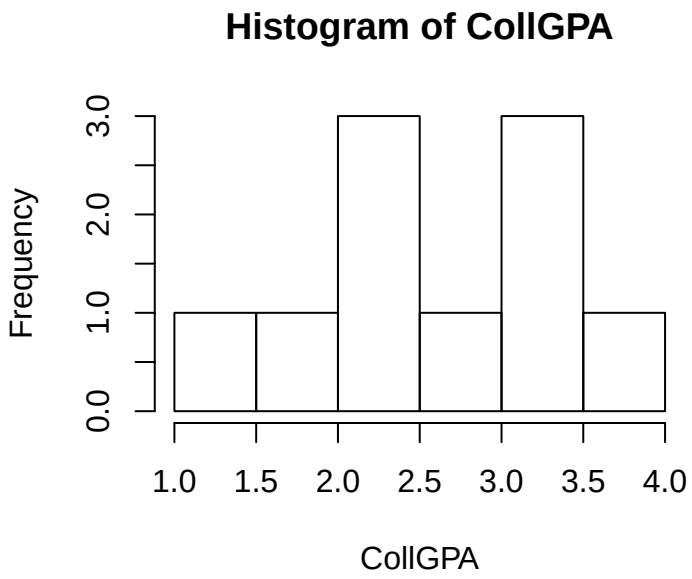
```
##      HSGPA          CollGPA
##  Min.   :2.000   Min.   :1.400
##  1st Qu.:2.425   1st Qu.:2.250
##  Median :2.650   Median :2.650
##  Mean   :2.710   Mean   :2.700
##  3rd Qu.:3.100   3rd Qu.:3.325
```

```
##  Max.   :3.400   Max.   :3.800
```

```
hist(HSGPA)
```



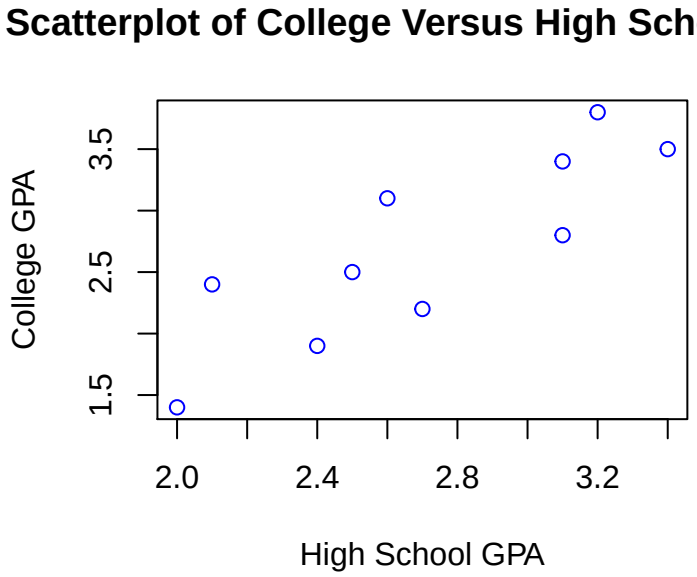

```
hist(CollGPA)
```




```
Y <- CollGPA
x <- HSGPA
```

Q3.1. Create a scatterplot of CollGPA versus HSGPA and find the correlation between teh two variables. What can we infer from the correlation?

```
plot(x, Y, col="blue", main="Scatterplot of College Versus High School GPA",
     xlab="High School GPA",ylab="College GPA")
```




```
cor(x,Y)
```

```
## [1] 0.8439231
```

Now that we know a few things about the data, we want to find a line that best represents the relationship between the variables. In other words, we want to draw a slope that comes closest to describing the data.

Q3.2. Characterize the equation mathematically. Find the least squares estimates of β_0 and β_1 using Equations $b_0 = \bar{y}\text{-}b_1\bar{x}$ and $b_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

```
b0 = "$\\beta_0$"
b1 = "$\\beta_1$"
```

CollGPA= β_1HSGPA + β_0

```
b1 <- sum( (x-mean(x))*(Y-mean(Y)) ) / sum( (x-mean(x))^2 )
b0 <- mean(Y)-b1*mean(x)
c(b0,b1)
```

```
## [1] -0.950366  1.346999
```

To perform the least square regression is R we can use the lm command. If you are interested use the help(lm) command to learn the different options for using this function. To relationship between the variables is defined in the lm command using a tilde ("~") between the vector containing the response variable and the vector containing the explanatory variable: lm(Y ~ x).

If you would like to know what else is stored in the variable you can use the attributes command: attributes().

Q3.3. Find the least squares estimates of β_0 and β_1 using the R function lm().

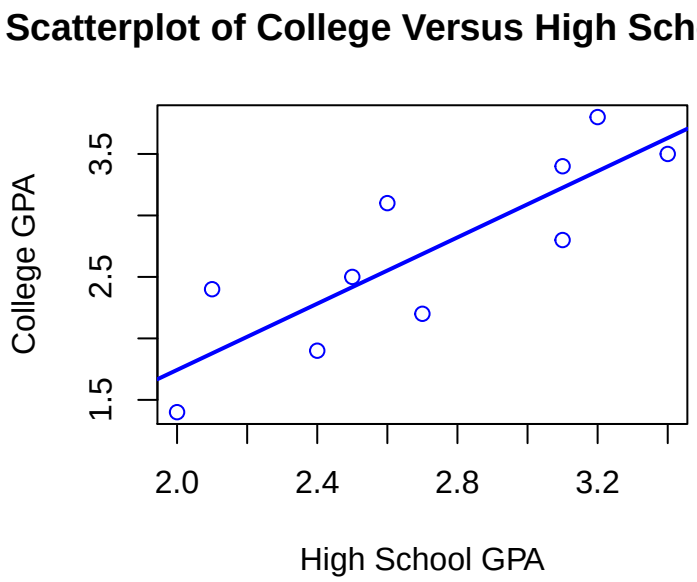
```
#model <- lm( CollGPA ~ HSGPA)
model <- lm(Y ~ x)
model$coef
```

```
## (Intercept)           x
##   -0.950366    1.346999
```

```
#coef(model)
```

Q3.4. abline() adds one or more straight lines to the current plot. The arguments to abline() are a=b0 and b=b1. Add the least squares line to the scatterplot created in 1 using the R function abline()

```
plot(x, Y, col="blue", main="Scatterplot of College Versus High School GPA",
     xlab="High School GPA",ylab="College GPA")
abline(model,col="blue",lwd=2)
```



OLS Goodness of Fit

When building regression models, "goodness-of-fit" explains how closely our model of the data (i.e. the predictor variables) fits the outcome data. In other words, how much of the variation in an outcome can we explain with a particular model?

R-Squared

R-squared is a measure commonly used for assessing model fit. It can be understood as the proportion of variance in the outcome that can be accounted for by the model.

Looking at our simple bivariate model, we can extract R-squared as a measure of model fit in a number of ways. The easiest is simply to extract it from the lm object using summary(model)$r.squared.

```
summary(model)$r.squared
```

```
## [1] 0.7122061
```

Warning: We normally discourage students from using the summary command with lm objects. The reason, as we will see later, is that summary makes a strong assumption called homoskedasticity, which is usually not justified. However, it is ok to use the command in order to extract R-squared.

But we can also calculate R-squared from our data in a number of ways. Take a couple of minutes to manually calculate R-squared.

1. By squaring the correlation between X and Y.
2. By taking the ratio of the variance of the fitted values to the variance of Y.
3. By weighting the slope coefficient: $R^2 = \beta_1^2 \frac{var(X)}{var(Y)}$

```r
cor(Y, x)^2  # as squared bivariate correlation
```

```
## [1] 0.7122061
```

```r
var(model$fitted)/var(Y)  # as ratio of variances
```

```
## [1] 0.7122061
```

```r
(coef(model)[2]/sqrt(cov(Y, Y)/cov(x, x)))^2  # as weighted regression coefficient
```

```
##         x
## 0.7122061
```

Adjusted R Square

The "Adjusted R-squared" is commonly used in place of the "regular" R-squared, which is sensitive to the number of independent variables in the model. In other words, as we put more variables into the model, R-squared increases even if those variables are unrelated to the outcome.

Adjusted R-squared attempts to correct for this by deflating R-squared by the expected amount of increase from including irrelevant additional predictors.

We can see this property of R-squared and Adjusted R-squared by adding a completely random variables unrelated to our other covariates or the outcome into our model and examine the impact on R-squared and Adjusted R-squared.

```r
tmp1 <- rnorm(10, 0, 10)
```

Add this variable to your simple regression model, creating a new lm object, then observe what happens to R-squared.

Now extract the adjusted R-squared from both models using lm$adj.r.squared. It may also go down, but by less than regular r-squared.

OLS: Issues to be Aware of

Unfortunately, the pitfalls of applying least squares are not often well understood by many of the people who attempt to apply it. What follows is a list of some of the biggest problems with using least squares regression in practice, along with some brief comments about how these problems may be mitigated or avoided

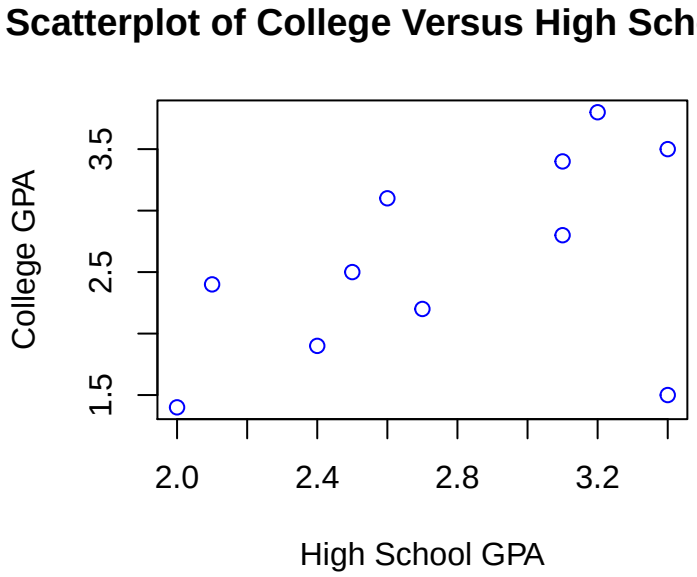
Outliers

Least squares regression can perform very badly when some points in the training data have excessively large or small values for the dependent variable compared to the rest of the training data. The reason for this is that since the least squares method is concerned with minimizing the sum of the squared error, any training point that has a dependent value that differs a lot from the rest of the data can have a disproportionately large effect on the resulting constants that are being solved for.

WARNING: Do not ever remove an observation just because it's an outlier.

Returning to our example, let's add an outlier.

```r
y_out <- c(Gpa$CollGPA,1.5)
x_out <- c(Gpa$HSGPA,3.4)
plot(x_out, y_out, col="blue", main="Scatterplot of College Versus High School GPA",
     xlab="High School GPA",ylab="College GPA")
```




```
cor(x_out,y_out)
```

```
## [1] 0.4986491
```

We can see the outlier pulls the correlation off a lot. Let's see what it does to the linear model.

```
model_out <- lm(y_out ~ x_out)
model_out$coef
```

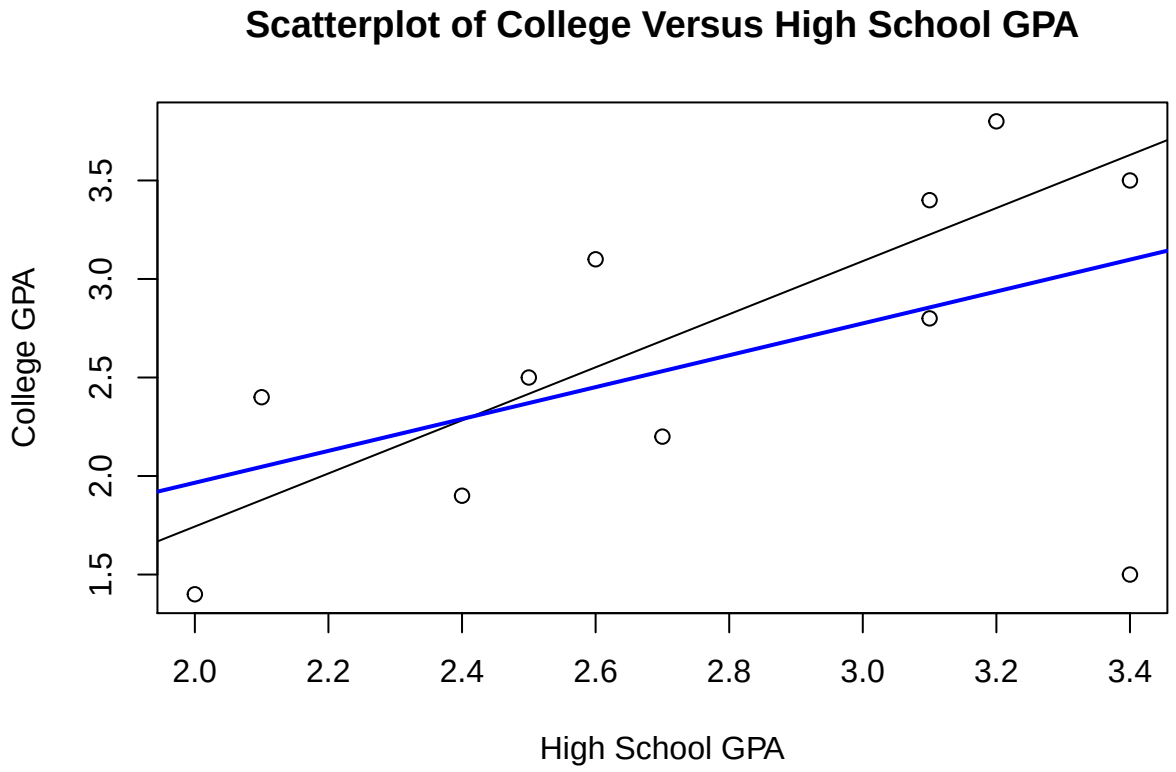
```
## (Intercept)       x_out
##   0.3483516   0.8087912
```

Let's see that scatterplot again with our new regression line.

```
attach(Gpa)
```

```
## The following objects are masked from Gpa (pos = 3):
##
##     CollGPA, HSGPA
```

```
regrline=(lm( CollGPA ~ HSGPA, data = Gpa))
plot(x_out, y_out, abline(regrline), main="Scatterplot of College Versus High School GPA",
     xlab="High School GPA",ylab="College GPA")
abline(model_out,col="blue",lwd=2)
```



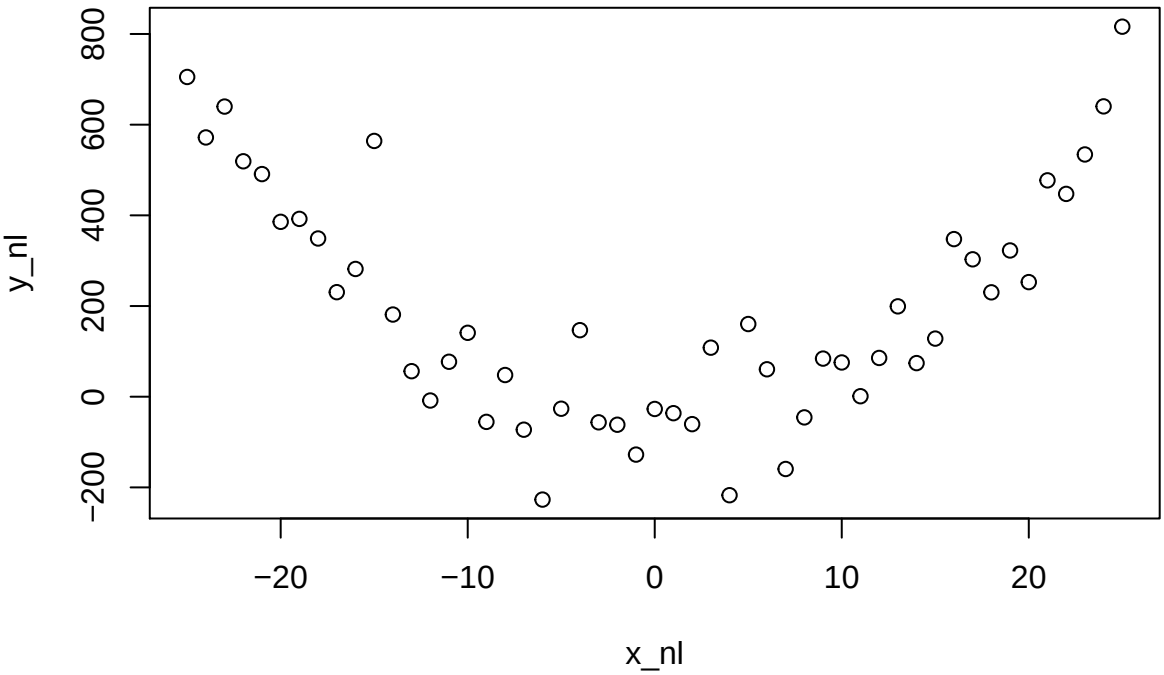
WARNING: Do not ever remove an observation just because it's an outlier.

Non-Linearities

All linear regression methods (including, of course, least squares regression), suffer from the major drawback that in reality most systems are not linear.

Let's take another dataset that is clearly non-linear.

```
x_nl<-seq(-25,25,1)
y_nl<-x_nl^2+rnorm(51,0,100)
plot(x_nl,y_nl)
```



There's definitely a relationship here, but we will need to do a transformation prior to OLS.