GV300 - Quantitative Political Analysis

University of Essex - Department of Government

Lorenzo Crippa

Week 11 – 9 December, 2019

Problem Set 4.1

1. (15 marks) Reviewing regression analysis basics: normal, t, χ^2 and F distributions. You will use your preferred statistical software for this exercise.

Problem Set 4.1

(a) (3 marks) generate a dataset with 2000 observations and create three variables each with values drawn from a standard normal distribution.

Problem Set 4.1

(a) (3 marks) generate a dataset with 2000 observations and create three variables each with values drawn from a standard normal distribution. For each variable **show that roughly 68% of the observations are within 1 standard deviation of 0, 95% within two standard deviations of 0, and 99% are within 3 standard deviations of zero.**

Problem Set 4.1

(a) (3 marks) generate a dataset with 2000 observations and create three variables each with values drawn from a standard normal distribution. For each variable **show that roughly 68% of the observations are within 1 standard deviation of 0, 95% within two standard deviations of 0, and 99% are within 3 standard deviations of zero.**

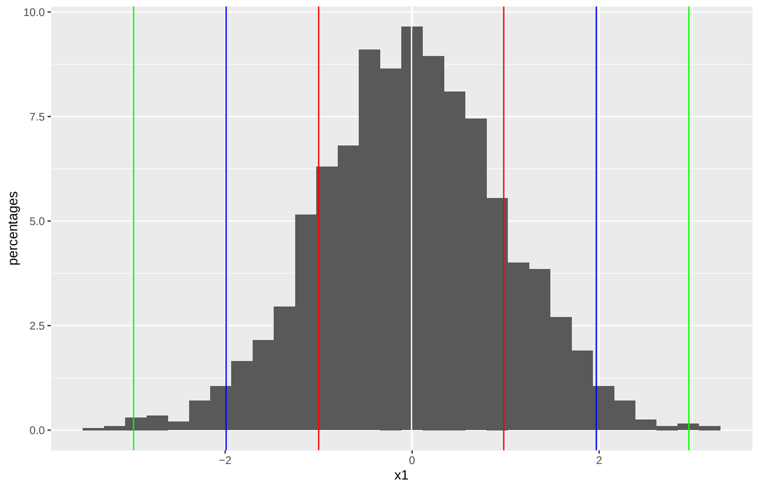
Show that roughly 68% of **your** observations are within 1 standard deviation of the mean, 95% within two standard deviations of the mean, and 99% are within 3 standard deviations.

Problem Set 4.1

(a) First determine (or represent) your three intervals.

Problem Set 4.1

(a) First determine (or represent) your three intervals. With a plot:



Problem Set 4.1

(a) Then use a percentile function.

Problem Set 4.1

(a) Then use a percentile function. In R:

Problem Set 4.1

(a) Then use a percentile function. In R:

```
quantile(df$x1, probs = c(.005, .025, .16, .84, .975,
    .99))
```

Problem Set 4.1

(a) Then use a percentile function. In R:

```
quantile(df$x1, probs = c(.005, .025, .16, .84, .975,
    .99))
```

In Stata:

```
centile x1, centile(.5 2.5 16 84 97.5 99)
```

(a) Then use a percentile function. In R:

```
quantile(df$x1, probs = c(.005, .025, .16, .84, .975,
    .99))
```

In Stata:

```
centile x1, centile(.5 2.5 16 84 97.5 99)
```

Output:

```
      0.5%       2.5%        16%        84% 
-2.7398407 -1.9486126 -0.9676747  0.9748144 

     97.5%        99% 
 1.8915366  2.2497764 
```

Problem Set 4.1

(a) Then use a percentile function. In R:

```
quantile(df$x1, probs = c(.005, .025, .16, .84, .975,
    .99))
```

In Stata:

```
centile x1, centile(.5 2.5 16 84 97.5 99)
```

Output:

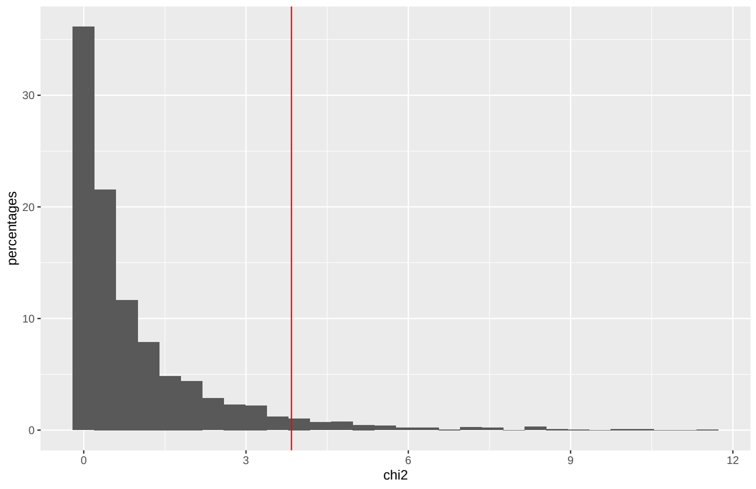
```
      0.5%        2.5%         16%         84%
-2.7398407 -1.9486126 -0.9676747  0.9748144

      97.5%        99%
  1.8915366  2.2497764
```

Repeat the same procedure for all variables

(b) (3 marks) **Show that 95% of the observations of a $\chi^2[1]$-distributed variable that you create are below the value that is associated with the 95$^{\text{th}}$ percentile of the $\chi^2[1]$-distribution.**

(b) (3 marks) **Show that 95% of the observations of a $\chi^2[1]$-distributed variable that you create are below the value that is associated with the 95^{th} percentile of the $\chi^2[1]$-distribution.** If you are unsure about how to do this, plot the $\chi^2[1]$-distribution and eye-ball which value is associated with the .95-percentile of that distribution.

(b) Distribution of a $\chi^2[1]$-distributed variable:



Problem Set 4.1

(b) From tables we know that 95^{th} percentile of a $\chi^2[1]$ distribution is 3.84

Problem Set 4.1

(b) From tables we know that 95^{th} percentile of a $\chi^2[1]$ distribution is 3.84 Again, we use a percentile function.

Problem Set 4.1

(b) From tables we know that 95th percentile of a $\chi^2[1]$ distribution is 3.84 Again, we use a percentile function.

In R:

Problem Set 4.1

(b) From tables we know that 95$^{\text{th}}$ percentile of a $\chi^2[1]$ distribution is 3.84 Again, we use a percentile function.

In R:

```
quantile(df$chi2, probs = .95)
```

Problem Set 4.1

(b) From tables we know that 95th percentile of a $\chi^2[1]$ distribution is 3.84 Again, we use a percentile function.

In R:

```
quantile(df$chi2, probs = .95)
```

In Stata:

```
centile chi2, centile(95)
```

(b) From tables we know that 95$^{\text{th}}$ percentile of a $\chi^2[1]$ distribution is 3.84 Again, we use a percentile function.

In R:

```
quantile(df$chi2, probs = .95)
```

In Stata:

```
centile chi2, centile(95)
```

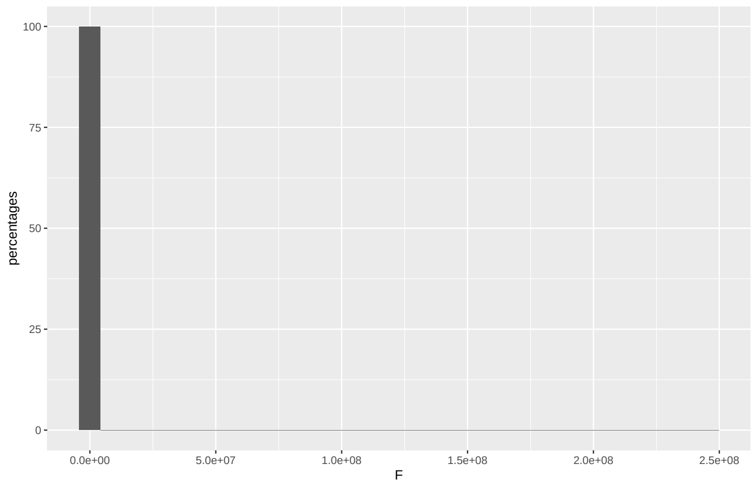
Output:

```
     95%
3.764813
```

Problem Set 4.1

(c) (3 marks) **Now, show that 95% of the observations of a $F[1,1]$ distributed variable that you create are below the value that is associated with the $.95$-percentile of the $F[1,1]$-distribution.**

(c) Distribution of an $F[1,1]$-distributed variable:



Problem Set 4.1

(c) From tables we know that 95^{th} percentile of a $F[1,1]$ distribution is 161.4

(c) From tables we know that 95^{th} percentile of a $F[1,1]$ distribution is 161.4 Again, we use a percentile function.

(c) From tables we know that 95^{th} percentile of a $F[1,1]$ distribution is 161.4 Again, we use a percentile function.

In R:

(c) From tables we know that 95^{th} percentile of a $F[1, 1]$ distribution is 161.4 Again, we use a percentile function.

In R:

```
quantile(df$F, probs = .95)
```

Problem Set 4.1

(c) From tables we know that 95^{th} percentile of a $F[1,1]$ distribution is 161.4 Again, we use a percentile function.

In R:

```
quantile(df$F, probs = .95)
```

In Stata:

```
centile F, centile(95)
```

Problem Set 4.1

(c) From tables we know that 95$^{\text{th}}$ percentile of a $F[1,1]$ distribution is 161.4 Again, we use a percentile function.

In R:

```
quantile(df$F, probs = .95)
```

In Stata:

```
centile F, centile(95)
```

Output:

```
     95%
168.6171
```

(d) (3 marks) Take one of the variables with a standard normal distribution from (a) and divide it by the square root of the $\chi^2[1]$-distributed variable you created in (b).

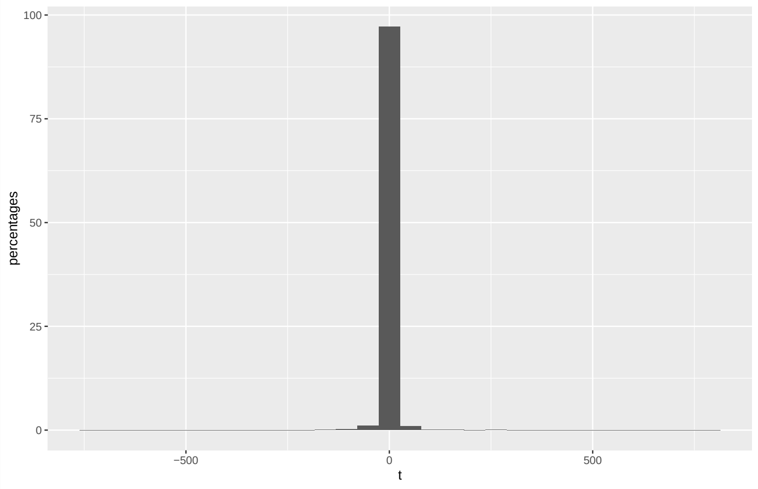
(d) (3 marks) Take one of the variables with a standard normal distribution from (a) and divide it by the square root of the $\chi^2[1]$-distributed variable you created in (b). Show that 95% of the observations of that new variable are below the value that is associated with the .95-percentile of the t-distribution.

(d) (3 marks) Take one of the variables with a standard normal distribution from (a) and divide it by the square root of the $\chi^2[1]$-distributed variable you created in (b). Show that 95% of the observations of that new variable are below the value that is associated with the .95-percentile of the t-distribution. For what purpose do we usually compute a t-statistic in regression analysis and how is it computed?

(d) (3 marks) Take one of the variables with a standard normal distribution from (a) and divide it by the square root of the $\chi^2[1]$-distributed variable you created in (b). Show that 95% of the observations of that new variable are below the value that is associated with the .95-percentile of the t-distribution. For what purpose do we usually compute a t-statistic in regression analysis and how is it computed? How does the computation of the t-statistic in regression analysis link to how you computed the new variable here in exercise (d)?

Problem Set 4.1

(d) Distribution of a $t[1]$-distributed variable:



(d) From tables we know that 95$^{\text{th}}$ percentile of a $t[1]$ distribution is 6.314

(d) From tables we know that 95$^{\text{th}}$ percentile of a $t[1]$ distribution is 6.314 Again, we use a percentile function.

Problem Set 4.1

(d) From tables we know that 95^{th} percentile of a $t[1]$ distribution is 6.314 Again, we use a percentile function.

In R:

(d) From tables we know that 95^{th} percentile of a $t[1]$ distribution is 6.314 Again, we use a percentile function.

In R:

```
quantile(df$t, probs = .95)
```

(d) From tables we know that 95^{th} percentile of a $t[1]$ distribution is 6.314 Again, we use a percentile function.

In R:

```
quantile(df$t, probs = .95)
```

In Stata:

```
centile t, centile(95)
```

(d) From tables we know that 95^{th} percentile of a $t[1]$ distribution is 6.314 Again, we use a percentile function.

In R:

```
quantile(df$t, probs = .95)
```

In Stata:

```
centile t, centile(95)
```

Output:

```
     95%
7.136391
```

(d) (3 marks) For what purpose do we usually compute a t-statistic in regression analysis and how is it computed?

(d) (3 marks) For what purpose do we usually compute a t-statistic in regression analysis and how is it computed? How does the computation of the t-statistic in regression analysis link to how you computed the new variable here?

We compute the t-statistic to conduct a hypothesis test about the size of the coefficient estimate.

(d) (3 marks) For what purpose do we usually compute a t-statistic in regression analysis and how is it computed? How does the computation of the t-statistic in regression analysis link to how you computed the new variable here?

We compute the t-statistic to conduct a hypothesis test about the size of the coefficient estimate.

The t-statistic is computed from the ratio of coefficient estimate minus reference value (most often zero, regression outputs report a zero reference value by default) and standard error of the estimate.

(d) (3 marks) For what purpose do we usually compute a t-statistic in regression analysis and how is it computed? How does the computation of the t-statistic in regression analysis link to how you computed the new variable here?

We compute the t-statistic to conduct a hypothesis test about the size of the coefficient estimate.

The t-statistic is computed from the ratio of coefficient estimate minus reference value (most often zero, regression outputs report a zero reference value by default) and standard error of the estimate.

The t-statistic is the ratio between OLS estimates (which have a standard normal distribution) and standard errors (squared roots of the variance, which has a χ^2 distribution)

(e) (3 marks) Plot all variables you created so far, that is, three variables with a standard normal distribution, one variable distributed $\chi^2[1]$, one variable distributed $F[1, 1]$, and one variable following the t-distribution.

Problem Set 4.1

(e) (3 marks) Plot all variables you created so far, that is, three variables with a standard normal distribution, one variable distributed $\chi^2[1]$, one variable distributed $F[1,1]$, and one variable following the t-distribution.

See previous graphs (histograms, but also densities ...)

2. (24 marks) Linking conditional expectation function and linear regression function. Load the dataset `baseball.csv`. It gives you information on a series of MLB players on their height in inches (variable *heightinches*) and weight (*weightpounds*).

Problem Set 4.2

(a) Generate the expected value of height for each value of weight.

Problem Set 4.2

(a) Generate the expected value of height for each value of weight.
It is simply the mean for values of weight (89 cases).

Problem Set 4.2

(a) Generate the expected value of height for each value of weight. It is simply the mean for values of weight (89 cases). In R:

Problem Set 4.2

(a) Generate the expected value of height for each value of weight. It is simply the mean for values of weight (89 cases). In R:

```
aggregate(data$heightinches, by = list(data$
    weightpounds), FUN = mean)
```

Problem Set 4.2

(a) Generate the expected value of height for each value of weight. It is simply the mean for values of weight (89 cases). In R:

```
aggregate(data$heightinches, by = list(data$
    weightpounds), FUN = mean)
```

In Stata:

```
tabstat heightinches, by(weightpounds)
```

Problem Set 4.2

(a) Generate the expected value of height for each value of weight. It is simply the mean for values of weight (89 cases). In R:

```
aggregate(data$heightinches, by = list(data$
    weightpounds), FUN = mean)
```

In Stata:

```
tabstat heightinches, by(weightpounds)
```

Output:

```
   Group.1          x
1      150 70.75000
2      155 69.33333
3      156 75.00000
4      160 71.46667
5      163 70.00000
...    ...
```

(b) (3 marks) Regress expected height on weight and record coefficient and standard error of that coefficient associated with the weight-variable. Interpret the outcome of this regression in words.

Problem Set 4.2

(b) (3 marks) Regress expected height on weight and record coefficient and standard error of that coefficient associated with the weight-variable. Interpret the outcome of this regression in words.

(c) (3 marks) Regress height on weight and record coefficient and standard error of that coefficient associated with the weight-variable. Interpret the outcome of this regression in words.

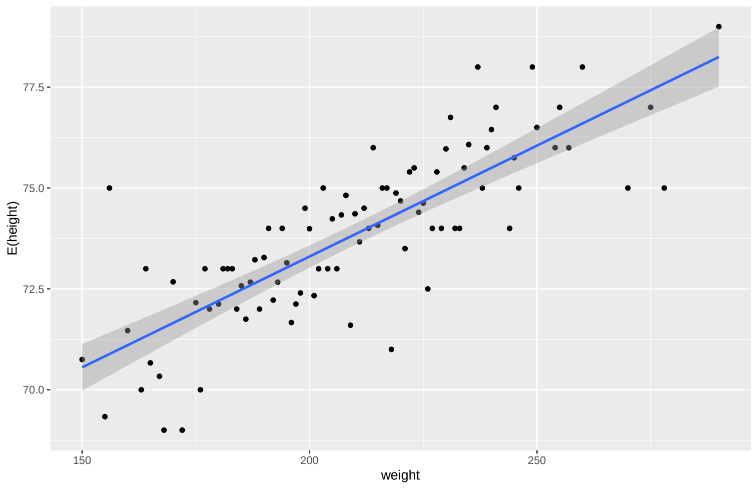
Problem Set 4.2

(b) (3 marks) Regress expected height on weight and record coefficient and standard error of that coefficient associated with the weight-variable. Interpret the outcome of this regression in words.

(c) (3 marks) Regress height on weight and record coefficient and standard error of that coefficient associated with the weight-variable. Interpret the outcome of this regression in words.

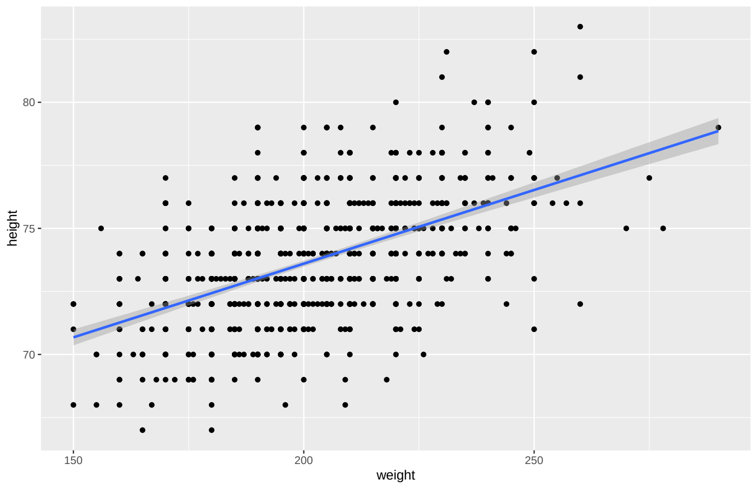
(d) (3 marks) Compare coefficients and standard errors in (b) and (c). What do you see?

(b) Represent:



Problem Set 4.2

(c) Represent:



Problem Set 4.2

(b), (c), (d)

Problem Set 4.2

(b), (c), (d) In R:

```
mod1 <- lm(data = df, expected.height ~ weightpounds)
mod2 <- lm(data = data, heightinches ~ weightpounds)

stargazer(mod1, mod2,keep.stat = c("n", "adj.rsq", "f"), type = "text")
```

Problem Set 4.2

(b), (c), (d) In R:

```
mod1 <- lm(data = df, expected.height ~ weightpounds)
mod2 <- lm(data = data, heightinches ~ weightpounds)

stargazer(mod1, mod2,keep.stat = c("n", "adj.rsq", "f"
    ), type = "text")
```

In Stata:

```
egen id = group(weightpounds)
reg exp_height weightpounds if id[_n] != id[_n+1]
est store mod1

reg heightinches weightpounds
est store mod2

esttab mod1 mod2, star(* .1 ** .05 *** .01) ar2
    scalars(F p) se
```

Problem Set 4.2

(b), (c), (d) Results:

(b), (c), (d) Results:

	Dependent variable:	
	E(height)	height
	(1)	(2)
weight	0.055***	0.058***
	(0.004)	(0.003)
Constant	62.315***	61.913***
	(0.918)	(0.588)
Observations	89	1,033
Adjusted R^2	0.645	0.282
F Statistic	161.183*** (df = 1; 87)	406.740*** (df = 1; 1031)
Note:		*p<0.1; **p<0.05; ***p<0.01

(e) (3 marks) Using the estimates of the regression of height on weight, what is the predicted height of someone who weights 225 pounds? If you do not know how to get a prediction out of your preferred statistical software, just compute by hand.

(e) (3 marks) Using the estimates of the regression of height on weight, what is the predicted height of someone who weights 225 pounds? If you do not know how to get a prediction out of your preferred statistical software, just compute by hand.

$$height = 61.913 + 0.058 weight = 61.913 + 0.058 \times 225 = 74.963$$

(f) 3 marks) Using the estimates of the regression of height on weight, what is the predicted height of someone who weighs 270 pounds?

(f) 3 marks) Using the estimates of the regression of height on weight, what is the predicted height of someone who weighs 270 pounds?

$$E(height) = 62.315 + 0.055 weight = 62.315 + 0.055 \times 270 = 77.165$$

Problem Set 4.2

(g) (3 marks) Using the estimates of the regression of height on weight and if you know that So Taguchi put on 30 pounds over the winter, how much do you predict his height changed?

Problem Set 4.2

(g) (3 marks) Using the estimates of the regression of height on weight and if you know that So Taguchi put on 30 pounds over the winter, how much do you predict his height changed?

It's a *ceteris paribus* question. Simply multiply the coefficient on weight by 30 to have the *marginal* change in height.

Problem Set 4.2

(g) (3 marks) Using the estimates of the regression of height on weight and if you know that So Taguchi put on 30 pounds over the winter, how much do you predict his height changed?

It's a *ceteris paribus* question. Simply multiply the coefficient on weight by 30 to have the *marginal* change in height. He should have grown by 1.74 inches.

(h) (3 marks) What do your answers to (e), (f), and (g) tell you about the interpretation of regression coefficients in general?

(h) (3 marks) What do your answers to (e), (f), and (g) tell you about the interpretation of regression coefficients in general?

It gives us a correlation, the direction of the effect is up to interpretation and not just the numbers.

3. (24 marks) Now, let's learn how to interpret all of the regression output more thoroughly. Input the data on campaign spending in a US Senatorial election below into your preferred statistical software:

3. (24 marks) Now, let's learn how to interpret all of the regression output more thoroughly. Input the data on campaign spending in a US Senatorial election below into your preferred statistical software:

District	Incumbent	Money	Vote Share
1	Matt Salmon	362	65
2	Ed Pastor	418	68
3	Jim Kolbe	712	52
4	Bob Stump	346	65
5	John Shadegg	426	69
6	J.D. Hayworth	1839	53

(a) (3 marks) Let's define the correlation between variables **M**oney and **V**ote Share as

$$cor(M, V) = \frac{cov(M, V)}{\sigma_M \sigma_V}$$

where the covariance of M and V is given by $cov(M, V) = 1/n \sum_{i=1}^{n} (m_i - \overline{m})(v_i - \overline{v})$. Compute $cor(M, V)$ by hand and show your computations. Interpret the result of your computation. Are you surprised by the result? Why?

Problem Set 4.3

$\overline{M} = 683.83$ $\overline{V} = 62$

Problem Set 4.3

$\overline{M} = 683.83$ $\overline{V} = 62$
$cov(M, V) = \frac{1}{6}(((362 - \overline{M})(65 - \overline{V}) + (418 - \overline{M})(68 - \overline{V}) + (712 - \overline{M})(52 - \overline{V}) + (346 - \overline{M})(65 - \overline{V}) + (426 - \overline{M})(69 - \overline{V}) + (1839 - \overline{M})(53 - \overline{V})) = -2676.17$

Problem Set 4.3

$\overline{M} = 683.83$ $\overline{V} = 62$
$cov(M, V) = \frac{1}{6}(((362 - \overline{M})(65 - \overline{V}) + (418 - \overline{M})(68 - \overline{V}) + (712 - \overline{M})(52 - \overline{V}) + (346 - \overline{M})(65 - \overline{V}) + (426 - \overline{M})(69 - \overline{V}) + (1839 - \overline{M})(53 - \overline{V})) = -2676.17$
$\sigma_M = 581.39$ $\sigma_V = 7.54$

Problem Set 4.3

$\overline{M} = 683.83$ $\overline{V} = 62$
$cov(M, V) = \frac{1}{6}(((362 - \overline{M})(65 - \overline{V}) + (418 - \overline{M})(68 - \overline{V}) + (712 - \overline{M})(52 - \overline{V}) + (346 - \overline{M})(65 - \overline{V}) + (426 - \overline{M})(69 - \overline{V}) + (1839 - \overline{M})(53 - \overline{V})) = -2676.17$
$\sigma_M = 581.39$ $\sigma_V = 7.54$
$\rho_{M,V} = \frac{-2676.17}{(581.39)(7.54)} = -0.61$

Problem Set 4.3

$\overline{M} = 683.83$ $\overline{V} = 62$
$cov(M, V) = \frac{1}{6}(((362 - \overline{M})(65 - \overline{V}) + (418 - \overline{M})(68 - \overline{V}) + (712 - \overline{M})(52 - \overline{V}) + (346 - \overline{M})(65 - \overline{V}) + (426 - \overline{M})(69 - \overline{V}) + (1839 - \overline{M})(53 - \overline{V})) = -2676.17$
$\sigma_M = 581.39$ $\sigma_V = 7.54$
$\rho_{M,V} = \frac{-2676.17}{(581.39)(7.54)} = -0.61$
There is a negative correlation between money and vote share in this senatorial election! That's interesting. Shouldn't we expect to see candidates who spend more money on their campaign get more votes?

(b) (8 marks) Using your preferred statistical software, run a linear regression of V on M.

(b) (8 marks) Using your preferred statistical software, run a linear regression of V on M. Plot V over M as well as the predicted values of vote share ($\hat{V}$) over M. Then add the regression line and horizontal line that indicate the residual.

Problem Set 4.3

(b) (8 marks) Using your preferred statistical software, run a linear regression of V on M. Plot V over M as well as the predicted values of vote share ($\hat{V}$) over M. Then add the regression line and horizontal line that indicate the residual.

In R:

```
mod <- lm(data = data, vote.share ~ money)
summary(mod)
```

Problem Set 4.3

(b) (8 marks) Using your preferred statistical software, run a linear regression of V on M. Plot V over M as well as the predicted values of vote share ($\hat{V}$) over M. Then add the regression line and horizontal line that indicate the residual.

In R:

```
mod <- lm(data = data, vote.share ~ money)
summary(mod)
```

In Stata:

```
reg vote_share money
```

Problem Set 4.3

(b) Results:

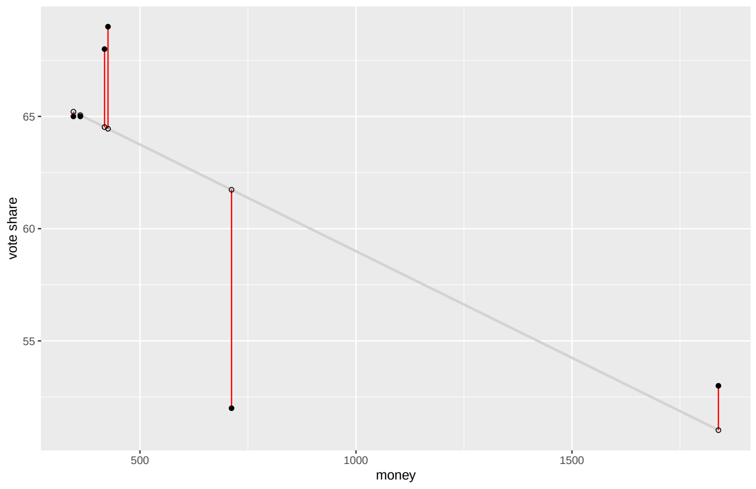
Problem Set 4.3

(b) Results:

	Dependent variable:
	vote.share
money	-0.010^{*}
	(0.004)
Constant	68.497^{***}
	(3.817)
Observations	6
Adjusted R^2	0.421
F Statistic	4.642^{*} (df = 1; 4)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Problem Set 4.3

(b) Plot:



(b) When you look at the coefficient estimate β_M, how is that estimate similar to the correlation between V and M you computed in 2(a)?

Problem Set 4.3

(b) When you look at the coefficient estimate β_M, how is that estimate similar to the correlation between V and M you computed in 2(a)? Are you surprised by the result?

Problem Set 4.3

(b) When you look at the coefficient estimate β_M, how is that estimate similar to the correlation between V and M you computed in 2(a)? Are you surprised by the result? Speculate about reasons why we see the relationship we see.

(b) When you look at the coefficient estimate β_M, how is that estimate similar to the correlation between V and M you computed in 2(a)? Are you surprised by the result? Speculate about reasons why we see the relationship we see.

Both statistics are negative, that is, the indicate a negative relationship between M and V.

(b) When you look at the coefficient estimate β_M, how is that estimate similar to the correlation between V and M you computed in 2(a)? Are you surprised by the result? Speculate about reasons why we see the relationship we see.

Both statistics are negative, that is, the indicate a negative relationship between M and V.

You should be surprised, again, we would have expected that candidates who spend more money generate higher vote shares.

(b) When you look at the coefficient estimate β_M, how is that estimate similar to the correlation between V and M you computed in 2(a)? Are you surprised by the result? Speculate about reasons why we see the relationship we see.

Both statistics are negative, that is, the indicate a negative relationship between M and V.

You should be surprised, again, we would have expected that candidates who spend more money generate higher vote shares.

Those candidates who won in close races (vote share closer to 50) have to spend more on their campaign because a close race means a strong challenger they had to beat (and outspend).

(b) When you look at the coefficient estimate β_M, how is that estimate similar to the correlation between V and M you computed in 2(a)? Are you surprised by the result? Speculate about reasons why we see the relationship we see.

Both statistics are negative, that is, the indicate a negative relationship between M and V.

You should be surprised, again, we would have expected that candidates who spend more money generate higher vote shares.

Those candidates who won in close races (vote share closer to 50) have to spend more on their campaign because a close race means a strong challenger they had to beat (and outspend). Expectations can thus reverse causality!

Problem Set 4.3

(b) When you look at the coefficient estimate β_M, how is that estimate similar to the correlation between V and M you computed in 2(a)? Are you surprised by the result? Speculate about reasons why we see the relationship we see.

Both statistics are negative, that is, the indicate a negative relationship between M and V.

You should be surprised, again, we would have expected that candidates who spend more money generate higher vote shares.

Those candidates who won in close races (vote share closer to 50) have to spend more on their campaign because a close race means a strong challenger they had to beat (and outspend). Expectations can thus reverse causality! Moreover, the sample might be very selected (6 observations)

Problem Set 4.3

(c) (3 marks) Can we reject the null hypothesis that there is no relationship between V and M?

Problem Set 4.3

(c) (3 marks) Can we reject the null hypothesis that there is no relationship between V and M?

The test distribution is the t-distribution and the t statistic in this example is -2.155.

Problem Set 4.3

(c) (3 marks) Can we reject the null hypothesis that there is no relationship between V and M?

The test distribution is the t-distribution and the t statistic in this example is -2.155.
When we apply the level of significance $\alpha = .05$, we cannot reject the null hypothesis that $\beta_1 = 0$.

(c) (3 marks) Can we reject the null hypothesis that there is no relationship between V and M?

The test distribution is the t-distribution and the t statistic in this example is -2.155.
When we apply the level of significance $\alpha = .05$, we cannot reject the null hypothesis that $\beta_1 = 0$.
The p-value is .0975.

Problem Set 4.3

(c) (3 marks) Can we reject the null hypothesis that there is no relationship between V and M?

The test distribution is the t-distribution and the t statistic in this example is -2.155.
When we apply the level of significance $\alpha = .05$, we cannot reject the null hypothesis that $\beta_1 = 0$.
The p-value is .0975.
The critical values of the t-distribution for $\alpha = .05$ for a two-sided test with 4 degrees of freedom are ± 2.78.

(c) (3 marks) Can we reject the null hypothesis that there is no relationship between V and M?

The test distribution is the t-distribution and the t statistic in this example is -2.155.
When we apply the level of significance $\alpha = .05$, we cannot reject the null hypothesis that $\beta_1 = 0$.
The p-value is .0975.
The critical values of the t-distribution for $\alpha = .05$ for a two-sided test with 4 degrees of freedom are ± 2.78.
The computed t statistic in our sample is not larger than that critical value.

(d) (4 marks) Now, run a regression of V on the intercept only. Show your results. What does the coefficient estimate represent?

(d) (4 marks) Now, run a regression of V on the intercept only. Show your results. What does the coefficient estimate represent?
Generate a new variable "m.low" which takes on value 1 if $M < 500$ and 0 otherwise. Run a regression of V on m.low.

(d) (4 marks) Now, run a regression of V on the intercept only. Show your results. What does the coefficient estimate represent?
Generate a new variable "m.low" which takes on value 1 if $M < 500$ and 0 otherwise. Run a regression of V on m.low. Compute the group-wise means of V of incumbents with low campaign spending *vs* those with high campaign spending from the regression results.

(d) (4 marks) Now, run a regression of V on the intercept only. Show your results. What does the coefficient estimate represent?
Generate a new variable "m.low" which takes on value 1 if $M < 500$ and 0 otherwise. Run a regression of V on m.low. Compute the group-wise means of V of incumbents with low campaign spending *vs* those with high campaign spending from the regression results. Show your computation.

(d) The coefficient simply represents the mean.

(d) The coefficient simply represents the mean.

In R:

```
mod2 <- lm(data = data, vote.share ~ vote.share)
summary(mod2)
mean(data$vote.share)
```

Problem Set 4.3

(d) The coefficient simply represents the mean.

In R:

```
mod2 <- lm(data = data, vote.share ~ vote.share)
summary(mod2)
mean(data$vote.share)
```

In Stata:

```
reg vote_share
sum vote_share
```

(d) Ordinal variable.

(d) Ordinal variable.

In R:

```
# ordinal variable
data$m.low <- ifelse(data$money < 500, 1, 0)

# model
mod3 <- lm(data = data, vote.share ~ m.low)
summary(mod3)

# group-wise mean
aggregate(data$vote.share, by = list(data$m.low), FUN = mean)
```

Problem Set 4.3

(d) Ordinal variable.

Problem Set 4.3

(d) Ordinal variable.

In Stata:

```
* ordinal variable
gen m_low = 1 * (money < 500)

* model
reg vote_share m_low

* group-wise mean
tabstat vote_share, by(m_low)
```

(e) (6 marks) Compute, by hand, the sum of squared residuals (SSR), the explained sum of squares (ESS), the total sum of squares (TSS), and R^2 for the regression of V on M.

(e) (6 marks) Compute, by hand, the sum of squared residuals (SSR), the explained sum of squares (ESS), the total sum of squares (TSS), and R^2 for the regression of V on M. Explain what R^2 tells you about model fit for this particular regression.

Problem Set 4.3

(e) (6 marks) Compute, by hand, the sum of squared residuals (SSR), the explained sum of squares (ESS), the total sum of squares (TSS), and R^2 for the regression of V on M. Explain what R^2 tells you about model fit for this particular regression.

$$TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

Problem Set 4.3

(e) (6 marks) Compute, by hand, the sum of squared residuals (SSR), the explained sum of squares (ESS), the total sum of squares (TSS), and R^2 for the regression of V on M. Explain what R^2 tells you about model fit for this particular regression.

$$TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2 \quad ESS = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$$

Problem Set 4.3

(e) (6 marks) Compute, by hand, the sum of squared residuals (SSR), the explained sum of squares (ESS), the total sum of squares (TSS), and R^2 for the regression of V on M. Explain what R^2 tells you about model fit for this particular regression.

$$TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2 \ ESS = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \text{ and } SSR = \sum_{i=1}^{n} \hat{u}^2$$

Problem Set 4.3

(e) (6 marks) Compute, by hand, the sum of squared residuals (SSR), the explained sum of squares (ESS), the total sum of squares (TSS), and R^2 for the regression of V on M. Explain what R^2 tells you about model fit for this particular regression.

$$TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2 \; ESS = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \text{ and } SSR = \sum_{i=1}^{n} \hat{u}^2$$

$TSS = 284$, $ESS = 152.56$, and $SSR = 131.44$.

Problem Set 4.3

(e) (6 marks) Compute, by hand, the sum of squared residuals (SSR), the explained sum of squares (ESS), the total sum of squares (TSS), and R^2 for the regression of V on M. Explain what R^2 tells you about model fit for this particular regression.

$TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2$ $ESS = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$ and $SSR = \sum_{i=1}^{n} \hat{u}^2$

$TSS = 284$, $ESS = 152.56$, and $SSR = 131.44$.

$R^2 = \frac{ESS}{TSS} = \frac{152.56}{284} = .5372$ (or $1 - SSR/TSS$), gives the explained variance.

4. (20 marks) Review how the ordinary least squares estimator is derived
   (a) (10 marks) Derive the ordinary least squares estimator β_1 and its sampling distribution for the population model

$$y = b_0 + b_1 x + e.$$

   Show every step of your derivation.

Problem Set 4.4

4. (20 marks) Review how the ordinary least squares estimator is derived
   (a) (10 marks) Derive the ordinary least squares estimator β_1 and its sampling distribution for the population model

   $$y = b_0 + b_1 x + e.$$

   Show every step of your derivation.
   (b) (10 marks) At which steps in your derivation did you make use of the six assumptions discussed in class (week 9 slides) and the text book? Clearly indicate where you made an assumption and explain in your own words what each assumption implies.

Problem Set 4.4

4. (20 marks) Review how the ordinary least squares estimator is derived
   (a) (10 marks) Derive the ordinary least squares estimator β_1 and its sampling distribution for the population model

   $$y = b_0 + b_1 x + e.$$

   Show every step of your derivation.
   (b) (10 marks) At which steps in your derivation did you make use of the six assumptions discussed in class (week 9 slides) and the text book? Clearly indicate where you made an assumption and explain in your own words what each assumption implies.

On the whiteboard

Problem Set 4.5

5. (17 marks) Finally, let's see why it is hard to get a causal claim out of regression analysis:
   (a) (5 marks) Generate a 2000 observation dataset.

Problem Set 4.5

5. (17 marks) Finally, let's see why it is hard to get a causal claim out of regression analysis:
   (a) (5 marks) Generate a 2000 observation dataset. Generate a variable "university" that equals 0 for the first 1000 observations and 1 for the second 1000 observations. This will represent half of the sample attending university.

5. (17 marks) Finally, let's see why it is hard to get a causal claim out of regression analysis:
   (a) (5 marks) Generate a 2000 observation dataset. Generate a variable "university" that equals 0 for the first 1000 observations and 1 for the second 1000 observations. This will represent half of the sample attending university. Generate a variable "income" which represents peoples' incomes. Let income = 15,000 + 5,000*university + 1,000*noise where "noise" is distributed standard normal.

5. (17 marks) Finally, let's see why it is hard to get a causal claim out of regression analysis:
   (a) (5 marks) Generate a 2000 observation dataset. Generate a variable "university" that equals 0 for the first 1000 observations and 1 for the second 1000 observations. This will represent half of the sample attending university. Generate a variable "income" which represents peoples' incomes. Let income = 15,000 + 5,000*university + 1,000*noise where "noise" is distributed standard normal.
   Regress income on university and show the regression output.

Problem Set 4.5

Obtaining variables and model

Problem Set 4.5

Obtaining variables and model

In R:

```r
df <- data.frame(
  university = c(rep(0, 1000), rep(1, 1000)),
  noise = rnorm(2000)
)

df$income <- 15000 + 5000*df$university +
             1000*df$noise

mod <- lm(data = df, income ~ university)
summary(mod)
```

Problem Set 4.5

Obtaining variables and model

Problem Set 4.5

Obtaining variables and model

In Stata:

```
clear
set obs 2000

gen university = 0 if _n <= 1000
replace university = 1 if _n > 1000

gen noise = rnormal()
gen income = 15000 + 5000 * university + 1000 * noise

reg income university
```

Problem Set 4.5

Results:

	Dependent variable:
	income
university	4,979.440***
	(44.989)
Constant	15,013.000***
	(31.812)
Observations	2,000
Adjusted R^2	0.860
F Statistic	12,250.400*** (df = 1; 1998)
Note:	*p<0.1; **p<0.05; ***p<0.01

Problem Set 4.5

(a) What is the coefficient estimate on university?

Problem Set 4.5

(a) What is the coefficient estimate on university? Why should you have known before you even ran the regression what the coefficient estimate approximately will be?

Problem Set 4.5

(a) What is the coefficient estimate on university? Why should you have known before you even ran the regression what the coefficient estimate approximately will be? Is this a causal effect?

Problem Set 4.5

(a) What is the coefficient estimate on university? Why should you have known before you even ran the regression what the coefficient estimate approximately will be? Is this a causal effect?

The effect of $+1$ university on income is $+5000$ because is what we start from in defining income itself. It is a causal effect because the 0-conditional mean assumption is met.

Problem Set 4.5

(b) (5 marks) We now further assume that education has **no** effect on earnings, but that smart people tend to both go to university and earn more money.

Problem Set 4.5

(b) (5 marks) We now further assume that education has **no** effect on earnings, but that smart people tend to both go to university and earn more money.
Clear your dataset and generate a new 2000 observation dataset. Generate 2 variables with uniform distributions between 0 and 1, called "intelligence" and "luck." Generate a variable "university" which equals 1 if intelligence+luck>1 and 0 otherwise.

(b) (5 marks) We now further assume that education has **no** effect on earnings, but that smart people tend to both go to university and earn more money.
Clear your dataset and generate a new 2000 observation dataset. Generate 2 variables with uniform distributions between 0 and 1, called “intelligence” and “luck.” Generate a variable “university” which equals 1 if intelligence+luck$>$1 and 0 otherwise.
Let income $=$ 15,000 $+$ 10,000*Intelligence $+$ 1,000*noise where “noise” is distributed standard normal.

(b) (5 marks) We now further assume that education has **no** effect on earnings, but that smart people tend to both go to university and earn more money.
Clear your dataset and generate a new 2000 observation dataset. Generate 2 variables with uniform distributions between 0 and 1, called "intelligence" and "luck." Generate a variable "university" which equals 1 if intelligence+luck>1 and 0 otherwise.
Let income = 15,000 + 10,000*Intelligence + 1,000*noise where "noise" is distributed standard normal.
Regress income on university. Show your the regression output. What's your coefficient estimate on university?

Problem Set 4.5

Obtaining variables and model

Problem Set 4.5

Obtaining variables and model

In R:

```
df <- data.frame(
  intelligence = runif(2000),
  luck = runif(2000),
  noise = rnorm(2000)
)

df$university <- ifelse(df$intelligence+df$luck>1, 1,
    0)
df$income <- 15000 + 10000*df$intelligence + 1000*df$
    noise

mod <- lm(data = df, income ~ university)
summary(mod)
```

Problem Set 4.5

Obtaining variables and model

Obtaining variables and model

In Stata:

```
clear
set obs 2000

gen intelligence = runiform()
gen luck = runiform()
gen noise = rnormal()

gen university = 1 * (intelligence + luck > 1)

gen income = 15000 + 10000 * intelligence + 1000 *
    noise

reg income university
```

Problem Set 4.5

Results:

	Dependent variable:
	income
university	3,391.231***
	(117.015)
Constant	18,331.420***
	(82.990)
Observations	2,000
Adjusted R^2	0.296
F Statistic	839.902*** (df = 1; 1998)
Note:	*p<0.1; **p<0.05; ***p<0.01

Problem Set 4.5

(c) (7 marks) Are the two regressions above different conceptually (that is with respect to how the regression enables us to learn something about the world)? Are the two regressions above different mechanically (that is with respect to how we try to get at an unbiased estimate of the true effect of university on income)?

Problem Set 4.5

1. Conceptually, the first represents clean experimental data (university is randomly assigned and normally distributed error term independent of university).

Problem Set 4.5

1. Conceptually, the first represents clean experimental data (university is randomly assigned and normally distributed error term independent of university). Causal effect of University on wage.

Problem Set 4.5

1. Conceptually, the first represents clean experimental data (university is randomly assigned and normally distributed error term independent of university). Causal effect of University on wage. In this case, we know the DGP

Problem Set 4.5

1. Conceptually, the first represents clean experimental data (university is randomly assigned and normally distributed error term independent of university). Causal effect of University on wage. In this case, we know the DGP
2. The second has university confounded by intelligence and luck.

Problem Set 4.5

1. Conceptually, the first represents clean experimental data (university is randomly assigned and normally distributed error term independent of university). Causal effect of University on wage. In this case, we know the DGP
2. The second has university confounded by intelligence and luck. OVB because luck affects our estimate of the effect of university on wage through the error term.

Problem Set 4.5

1. Conceptually, the first represents clean experimental data (university is randomly assigned and normally distributed error term independent of university). Causal effect of University on wage. In this case, we know the DGP
2. The second has university confounded by intelligence and luck. OVB because luck affects our estimate of the effect of university on wage through the error term. In this case, we do not know the DGP

Problem Set 4.5

1. Conceptually, the first represents clean experimental data (university is randomly assigned and normally distributed error term independent of university). Causal effect of University on wage. In this case, we know the DGP
2. The second has university confounded by intelligence and luck. OVB because luck affects our estimate of the effect of university on wage through the error term. In this case, we do not know the DGP
3. We should have modelled wages as function of intelligence **and** luck.

Problem Set 4.5

1. Conceptually, the first represents clean experimental data (university is randomly assigned and normally distributed error term independent of university). Causal effect of University on wage. In this case, we know the DGP
2. The second has university confounded by intelligence and luck. OVB because luck affects our estimate of the effect of university on wage through the error term. In this case, we do not know the DGP
3. We should have modelled wages as function of intelligence **and** luck.
4. Mechanically, in the second case we are trying to estimate a coefficient on university which is a compound of the effect of intelligence and luck.

Conclusion

All clear? Questions?

Thanks and see you next week!