This is a draft of the solutions to the exam that was held on 21-08-17.

All students get the same seven problems, regardless of anonymity code.

Answer form for multiple choice. You can make your own form, put please be clear and answer on one page. Do not submit solutions to the multiple-choice problems.

Number	Part	A	B	C	D	E
1	a.	■	□	□	□	□
1	b.	■	□	□	□	□
2	a.	□	■	□	□	□
2	b.	□	□	■	□	□
3	a.	□	□	□	□	■
3	b.	□	□	□	■	□
4	a.	□	□	□	■	□
4	b.	□	□	■	□	□
5	a.	□	□	■	□	□
5	b.	□	■	□	□	□

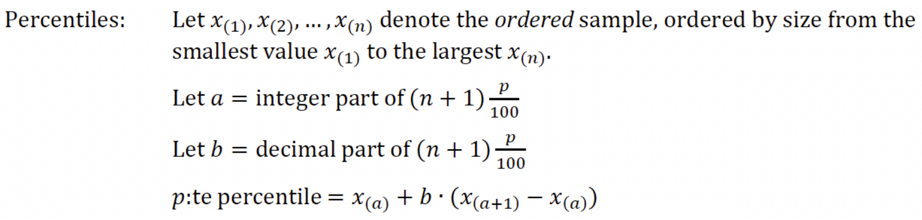
1.

a) ***Find the inter quartile range of the number of hours worked.***

The inter quartile range, or IQR, is the difference between the third and first quartiles, so we need to find those first. There are only 20 work times, so we can easily order and list all the values:

0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 2 3 4 4

We use the formula sheet to remind ourselves how percentiles work:



Percentiles: Let $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ denote the *ordered* sample, ordered by size from the smallest value $x_{(1)}$ to the largest $x_{(n)}$.

Let $a =$ integer part of $(n+1)\frac{p}{100}$

Let $b =$ decimal part of $(n+1)\frac{p}{100}$

p:te percentile $= x_{(a)} + b \cdot (x_{(a+1)} - x_{(a)})$

Then, calculate: $(n+1)\frac{p}{100} = (20+1)\frac{75}{100} = 15.75$

and: $(n+1)\frac{p}{100} = (20+1)\frac{25}{100} = 5.25$

This means that the third quartile is between the 15th and the 16th time in the list and that the first quartile is between the 5th and the 6th time in the list. Let us look at the list again:

0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 2 3 4 4

Here, the 5th, 6th, 15th, and 16th times are highlighted in green. We see that both the first and third quartiles are equal to 1, so the $IQR = 1 - 1 = 0$.
Answer: A

b) ***Find the sample standard deviation of the number hours studied.***

We have to equivalent variants of the formula for sample variance:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

Let us use the version on the right.
We have 4 zeros, 12 ones, 1 two, 1 three, and 2 fours.

$$\bar{x} = \frac{4\cdot 0 + 12\cdot 1 + 1\cdot 2 + 1\cdot 3 + 2\cdot 4}{20} = \frac{25}{20} = \frac{5}{4}.$$

$$\sum_{i=1}^{20} x_i^2 = 4\cdot 0^2 + 12\cdot 1^2 + 1\cdot 2^2 + 1\cdot 3^2 + 2\cdot 4^2 = 57$$

$$s_x^2 = \frac{\sum_{i=1}^{20} x_i^2 - n\bar{x}^2}{n-1} = \frac{57 - 20 \cdot \left(\frac{5}{4}\right)^2}{20-1}$$

$$s_x = \sqrt{\frac{57 - 20 \cdot \left(\frac{5}{4}\right)^2}{20-1}} = 1.1642$$

Answer: A

2.

a) ***If the car has a visible flaw or does not function properly (or both), the car will be discarded. Find the probability that the car will be discarded.***

Let *A* = "car has a visible flaw" and *B* = "car does not function properly." Then

$$P(A) = 0.02$$
$$P(B) = 0.01$$
$$P(A \cap B) = 0.002$$

and we seek $P(A \cup B)$. Some might think that the intersection should be 0.0002, but this is only true if *A* and *B* are independent, which they are not in this problem. Different flaws in production can certainly be dependent.

We need the

Addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$P(A \cup B) = 0.02 + 0.01 - 0.002 = 0.028$$

Answer: B

b) ***Find the probability that a randomly chosen car does not function properly, given that it has a visible flaw.***

Use the definition of conditional probability:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.002}{0.02} = 0.1$$

Answer: C

3.

a) ***Find the probability that at least four of the eleven scratch cards are winning tickets.***

We can view the eleven scratch cards as eleven experiments with fixed probability of win/success 0.25. They are also independent, so the number of wins follows a binomial distribution.

$$X \sim Binomial(n = 11, P = 0.25)$$

We seek "at least four" which is the complement of "three or fewer." So,

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - 0.71330 = 0.2876$$

The probability $P(X \leq 3)$ can be found on page 16 in the formula sheet.

Answer: E

b) ***Find the probability that Serhiy wins the race against William.***

Let Serhiy's finish time be X and Williams finish time be Y. Then,

$X \sim N(25, 1.5^2)$ and $Y \sim N(26, 2^2)$ with $\rho_{xy} = 0.2$

We seek $\mathrm{P}(\mathrm{X} < \mathrm{Y}) = \mathrm{P}(\mathrm{X} - \mathrm{Y} < 0)$. We need to standardize, so first we need the expected value and variance of the random variable $\mathrm{X} - \mathrm{Y}$. We will use these formulas, with $a = 1, b = -1, c = 0$:

$E(aX + bY + c) = aE(X) + bE(Y) + c$ $= a\mu_X + b\mu_Y + c$	$Var(aX + bY + c)$ $= a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$ $= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}$

We do not have the covariance, so we need to rearrange this formula:

Correlation between X and Y:

$$\rho_{XY} = Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}$$

$$\mathrm{Cov}(X,Y) = \rho_{xy}\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}$$

$$\begin{aligned}
&\mathrm{E}[X - Y] = \mathrm{E}[X] - \mathrm{E}[Y] = 25 - 26 = -1 \\
&\mathrm{Var}(X - Y) = \mathrm{Var}(1 \cdot X + (-1) \cdot Y + 0) = \\
&= 1^2 \cdot \mathrm{Var}(X) + (-1)^2 \cdot \mathrm{Var}(Y) + 2 \cdot 1 \cdot (-1) \cdot \rho_{xy}\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)} \\
&= 1^2 \cdot 1.5^2 + (-1)^2 \cdot 2^2 + 2 \cdot 1 \cdot (-1) \cdot 0.2\sqrt{1.5^2 \cdot 2^2} = 5.05
\end{aligned}$$

Now we are ready to standardize:

$$\mathrm{P}(\mathrm{X} - \mathrm{Y} < 0) = \mathrm{P}\left(\frac{X - Y - (-1)}{\sqrt{5.05}} < \frac{0 - (-1)}{\sqrt{5.05}}\right) = \mathrm{P}\left(Z < \frac{0 - (-1)}{\sqrt{5.05}}\right) \approx \mathrm{P}(Z < 0.44)$$

This is a probability that we can get straight from table 1:

$$\mathrm{P}(Z < 0.44) = F(0.44) = 0.67003$$

Answer: D

4.

a) ***Find a 99% confidence interval for the difference in mean male height between the two states, $\mu_{NJ} - \mu_{AZ}$.***

We have two independent samples. The sample sizes from the two populations are less than 30, and since we only have sample standard deviation, the sampe sizes are unknown. Hence, we will use:

σ_X^2, σ_Y^2 unknown and assumed equal:	$\bar{x} - \bar{y} \pm t_{n_x+n_y-2;\alpha/2} \cdot s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$ where $s_p^2 = \frac{(n_x-1)s_X^2+(n_y-1)s_Y^2}{n_x+n_y-2}$

We need

$$\begin{aligned}&\bar{x} = 177.4\\&\bar{y} = 176.2\\&\alpha = 1 - 0.99 = 0.01\\&\frac{\alpha}{2} = 0.005\\&n_x = n_y = 10\\&\nu = \ 10 + 10 - 2 = 18\\&t_{18;0.005} = 2.878\\&s_x = s_y = 5.1\end{aligned}$$

We can use the formula to find s_p^2 if we want, but when the sample sizes are equal, s_p^2 is just the mean of the variances, as discussed in class. So, $s_p^2 = 5.1^2$ and $s_p = 5.1$. Now we can put it all together:

$$177.4 - 176.2 \pm 2.878 \cdot 5.1 \cdot \sqrt{\frac{1}{10} + \frac{1}{10}} \Rightarrow (-5.36, 7.76)$$

Answer: D

b) ***Assuming that the sample is representative of the whole population, <u>find the margin of error</u> for the proportion of British adults who have given money to charitable organizations in the past 12 months. Use a 95% confidence interval.***

The formula for the confidence interval is:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The margin of error is half the width of the confidence interval, so it is just this part:

$$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\begin{aligned}&\hat{p} = \frac{445}{1000} = 0.445\\&\alpha = 1 - 0.95 = 0.05\\&\frac{\alpha}{2} = 0.025\\&z_{0.05} = 1.96\\&n = 1000\end{aligned}$$

So,

$$ME = 1.96\sqrt{\frac{0.445(1 - 0.445)}{1000}} = 0.031$$

Answer: C

5.

a) ***Find the critical value of the test.***

This is an independence test, so the variable is χ^2-distributed. How many degrees of freedom? We summarize the choices and genders in a table:

	blue	green
female	48	24
male	66	22

To get the degrees of freedom, we use the formula $(r-1)(c-1)$ where r and c are the number of rows and columns, so $\nu = (2-1)(2-1) = 1$. The level of significance is 1%, so we find 1% and 1 degree of freedom in table 4: $\chi^2_{crit} =$ 6.635.
Answer: C

b) ***Find the value of the test variable.***
The formula:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad \text{where } E_{ij} = \frac{R_i C_j}{n}$$

We make some tables. Observed counts, row and column sums:

	blue	green	Ri
female	48	24	72
male	66	22	88
Ci	114	46	160

Expected counts under the null, E_{ij}:

E	C1	C2
R1	51.3	20.7
R2	62.7	25.3

Difference between expected and observed, $O_{ij} - E_{ij}$:

D	C1	C2
R1	-3.3	3.3
R2	3.3	-3.3

And finally, we can calculate the terms $\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

D^2/E		
	0.2122807	0.52608696
	0.17368421	0.43043478

When we sum these four cells, we get precisely the sum in the formula and the answer: 1.34248665
Answer: B

6.

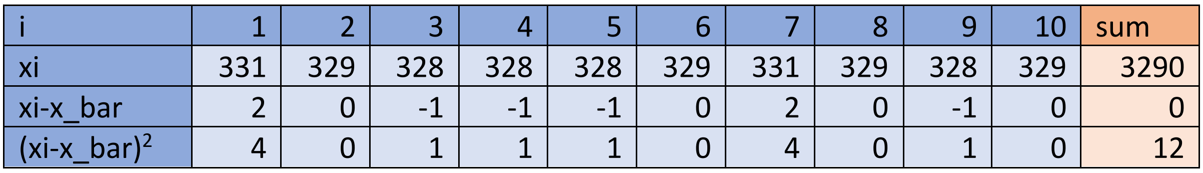
a) ***Find the sample mean and sample variance of the sample.***

First, calculate the mean of x; it is 329 m1.
We have to equivalent variants of the formula for sample variance:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

Let us use first version (on the left) this time.

Then we can create a table:



i	1	2	3	4	5	6	7	8	9	10	sum
xi	331	329	328	328	328	329	331	329	328	329	3290
xi-x_bar	2	0	-1	-1	-1	0	2	0	-1	0	0
(xi-x_bar)^2	4	0	1	1	1	0	4	0	1	0	12

The sum of the last row, 12, is the value of numerator of our formula. So,

$$s_x^2 = \frac{12}{10-1} = \frac{12}{9} = \frac{4}{3}$$

b) ***State necessary assumptions, hypotheses and the test variable.***

Assumptions: the sample is i.i.d. which means Independent and Identically Distributed. The other assumption is that the population is normally distributed, but this does not have to be mentioned since it is stated in the question text.

Hypotheses:

$$H_0: \mu = 330$$
$$H_1: \mu < 330$$

Test variable:
We have small sample size, normal distribution, and unknown variance, so:

σ_X^2 unknown: $t_{n-1} = \frac{\bar{X} - \mu_0}{s_X/\sqrt{n}}$

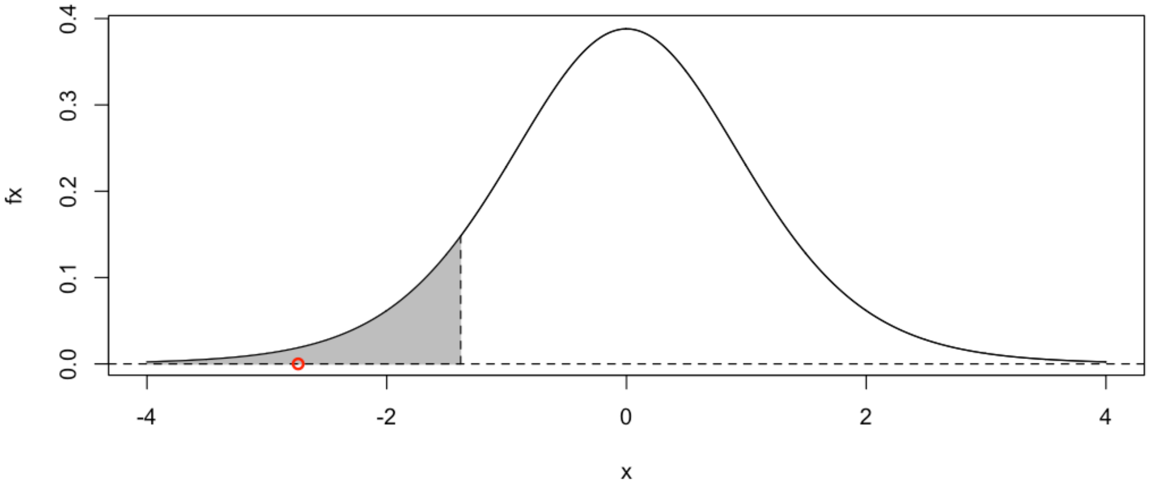
c) ***State the critical value and decision rule.***

Critical value:

$$\nu = n - 1 = 10 - 1 = 9$$
$$\alpha = 0.10$$
$$t_{0.1;9} = 1.383$$

Decision rule:

We reject the null if $t_{obs} < -1.383$



We want the critical region to be 10% of the area and to the left of zero, as illustrated above. The observed test statistic from (d) is marked in red.

d) ***Calculate the test statistic and draw conclusion.***

$$\bar{x} = 329$$
$$\mu_0 = 330$$
$$s_x^2 = \frac{4}{3}$$
$$s_x = \sqrt{\frac{4}{3}}$$
$$n = 10$$

$$t_{obs} = \frac{329 - 330}{\sqrt{4/3}\,/\,\sqrt{10}} = \frac{-1}{\sqrt{4/30}} = -\frac{\sqrt{30}}{2} = -2.738613$$

<u>Conclusion:</u> Since the observed test variable is -2.739, which <u>is</u> less than -1.383, we <u>reject</u> the null. We have found significant evidence, at the 10% level, for the hypothesis that the mean volume of soda is less than 330 ml.

e) ***Find the probability that someone guesses four out of four correctly, if they cannot tell the difference at all and they just guess.***

If you are asked to pick the four classes out of eight that contain White Lightning Soda, there is only one correct answer. For example, if you number the glasses 1 through 8, the correct answer might be 2, 3, 6, and 8. Does order matter? No. The answer 8, 3, 6, and 2 would be the same answer as 2, 3, 6, and 8.
Is this with or without replacement? You cannot choose the same class twice, so this is without replacement. The total number of possible different answers is given by the formula:

$$C_k^n = \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

(Without replacement, orderd does not matter)

$$\binom{8}{4} = \frac{8!}{4!\,(8-4)!} = \frac{8\cdot7\cdot6\cdot5\cdot\cancel{4\cdot3\cdot2\cdot1}}{\cancel{4\cdot3\cdot2\cdot1}\cdot4\cdot3\cdot2\cdot1} = \frac{\cancel{8}\cdot7\cdot6\cdot5}{4\cdot3\cdot\cancel{2}\cdot1} = \frac{7\cdot6\cdot5}{3\cdot1} = 7\cdot2\cdot5 = 70$$

There are 70 possible choices and 1 correct choice. Therefore, someone who guesses has a 1/70 probability of guessing correctly.

Alternative solution: Your first guess, you have 4/8 chance of picking a White Lightning glass. Then there are three White Lightning glasses left and seven glasses total. So, for your next pick, you have 3/7 chance of getting it right, and so on. This comes to:

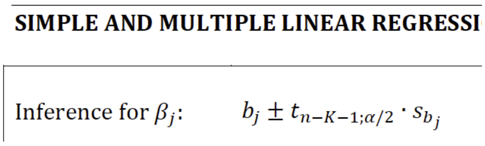
$$\frac{4}{\cancel{8}}\cdot\frac{3}{7}\cdot\frac{\cancel{2}}{6}\cdot\frac{1}{5} = \frac{3\cdot1}{7\cdot6\cdot5} = \frac{1}{7\cdot2\cdot5} = \frac{1}{70}$$

Note that this is about 1.4% chance, so if I witnessed Jane do this, I would believe that she really can taste the difference.

7.

(a) ***Find a 95% confidence interval for the variable BEDROOMS in model 3. Interpret the result.***

The formula:



SIMPLE AND MULTIPLE LINEAR REGRESSI

Inference for β_j: $b_j \pm t_{n-K-1;\alpha/2} \cdot s_{b_j}$

$$\begin{aligned}
&b_2 = 14.79\\
&n = 50\\
&K = 5\\
&n-K-1 = 46\\
&\alpha = 0.05\\
&\alpha/2 = 0.025\\
&t_{44;0.025} \approx t_{45;0.025} = 2.014\\
&s_{b_2} = \; 22.46\\
&UCL = \; 14.79 + 2.014\cdot22.46 = 60.02\\
&LCL = \; 14.79 - 2.014\cdot22.46 = -30.44
\end{aligned}$$

Interpretation: We can say with "95% confidence" that the expected price change of a home is between -30 thousand dollars and 60 thousand dollars, if we add one more bedroom, **with everything else held constant**. This means that if we compare homes with the same living area, latitude, longitude, and they are either both on the waterfront or not on the waterfront, then the home with one more bedroom is worth between -30,000 and 60,000 more than the home with one less bedroom. (This means that the slope is not significantly different from zero at the 5% level.)

For parts (b) and (c), you are asked to test whether the coefficient of the variable LONG is significantly different from zero, given that SQFT is included in the model. Use 5% level of significance.

(b) ***State the hypotheses, test variable, critical value, and decision rule.***

Hypotheses:

$$H_0: \beta_2 = 0 \mid \beta_1 \neq 0$$
$$H_1: \beta_2 \neq 0 \mid \beta_1 \neq 0$$

Test variable:

$$t_{n-K-1} = \frac{b_j - \beta_j^*}{s_{b_j}}$$

Critical value:

$$n = 50$$
$$K = 2$$
$$n - K - 1 = 47$$
$$\alpha = 0.05$$
$$\alpha/2 = 0.025$$
$$t_{47;0.025} \approx t_{45;0.025} = 2.014$$

Decision Rule: We reject the null if $|t_{obs}| > 2.014$.

(c) ***Calculate the value of the test statistic and state your conclusions.***

$$t_{obs} = \frac{-3607.4 - 0}{1572.5} = -2.29$$

Conclusion: since $|t_{obs}| = |-2.29| = 2.29 > 2.014$, we reject the null. We have found significant evidence at the 5% level for the hypothesis that the slope of *LONG* is different from zero. In other words, where in the how far east the house is located within the neighborhood affects the price of the home.

d) ***Use model 1 to find a 90% prediction interval for the price of a home given that the living area of the house is 1500 square feet. Interpret the result.***

Formula:

Prediction interval for the prediction of y given $X = x$:

$$(b_0 + b_1 x) \pm t_{n-2,\alpha/2} \sqrt{s_e^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2}\right)}$$

$$b_0 = -83.27$$
$$b_1 = 0.2713$$
$$x = 1500$$
$$n = 50$$
$$n - 2 = 48$$
$$\alpha = 0.10$$

$$\frac{\alpha}{2} = 0.05$$

$$t_{48;0.05} \approx t_{45;0.05} = 1.679$$

(You may round to 45 or 50 degrees of freedom.)

$$s_e^2 = 20802$$

(There are many ways to get this value. For example, you can take MSR straight from the table or take the square of “Standard Error”)

$$s_x^2 = 552284$$

$$\bar{x} = 1650$$

$$UCL = (-83.27 + 0.2713 \cdot 1500) + 1.679\sqrt{20802 \cdot \left(1 + \frac{1}{50} + \frac{(1500-1650)^2}{(50-1)552284}\right)}$$

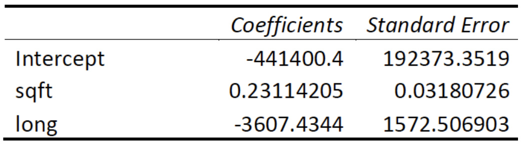
$$LCL = (-83.27 + 0.2713 \cdot 1500) - 1.679\sqrt{20802 \cdot \left(1 + \frac{1}{50} + \frac{(1500-1650)^2}{(50-1)552284}\right)}$$

This gives us: (79.0, 568.3)

Interpretation: According to Model 1, if we choose a random house in the neighborhood that has 1500 square feet living area, the price will be between \$79,000 and \$568,300 with 90% probability. There is a lot of variation in price for houses this size, in this neighborhood, according to the model.

e) *The correlation between the variables LONG and SQFT is negative (-0.55).* ***Use this information to briefly explain why the coefficient for SQFT is smaller in model 2, compared to model 1.***

This problem was meant to be a little tricky, but it illustrates an important concept in linear regression. According to the problem text, *LONG* gets bigger when we move east on the map. First, let us note that according to model 2, homes become less valuable when we move east:



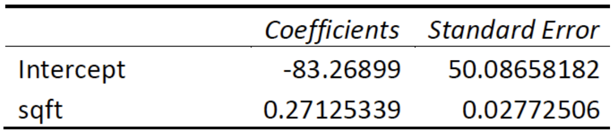
	Coefficients	*Standard Error*
Intercept	-441400.4	192373.3519
sqft	0.23114205	0.03180726
long	-3607.4344	1572.506903

That *LONG* is negatively correlated with *SQFT* means that when we move east, the average size of homes also becomes smaller. In other words, houses in the west are bigger on average than houses in the east.

MODEL 1: $PRICE = \beta_0 + \beta_1 * SQFT + \varepsilon$

MODEL 2: $PRICE = \beta_0 + \beta_1 * SQFT + \beta_2 * LONG + \varepsilon$

In model 1, we do not include east-west position in the model. If we look at the homes with larger living area (bigger homes), they are also on average further west. So, the coefficient *SQFT* in model one captures at least two things: the size of the homes and some of the east-west position of the homes.



	Coefficients	*Standard Error*
Intercept	-83.26899	50.08658182
sqft	0.27125339	0.02772506

In model 2, we control for east-west position, since the variable *LONG* is included in the model. In that model, the coefficient for *SQFT* is the change in price when we increase living area one square foot, **while holding east-west position constant**. Now an increase in living space no longer means that the house is further west, on average. Therefore, the coefficient is smaller than it is in model 1.

Of course, you would not have to write all of that to get full credit.