Spring 2020 Applied Statistics Comprehensive Examination Solutions

1. (15 points) Suppose that a television critic hypothesizes that the variation in run time per episode of popular cable television shows has been increasing in recent years. She decides to obtain two random samples of 30 episodes from all episodes shown among the 10 most popular "hour-long" cable television shows in 2009 and 2019. The following means and standard deviations are obtained for run times:

Year	n	$\overline{y}$	s
2009	30	51.3	2.1
2019	30	54.2	3.9

   (a) (10 points) Conduct a 0.05 level test to determine if the variance in run times is higher in 2019 than in 2009.

   Solution:
   $H_0 : \sigma^2_{2019} \leq \sigma^2_{2009}$ vs. $H_a : \sigma^2_{2019} > \sigma^2_{2009}$

   where σ^2_{2019} is the variance in popular cable television show run times in 2019 and σ^2_{2009} is the variance in popular cable television show run times in 2009

   F-stat $= \frac{3.9^2}{2.1^2} = 3.449$

   Critical/Rejection Region: F-stat $> F_{29,29,0.05} = 1.86 \approx F_{30,29,0.05} = 1.85$

   Then, since our F-statistic is in the rejection region, we have enough evidence at the 0.05 level to conclude that the variance in run times for popular television shows was higher in 2019 than in 2009.

   (b) (5 points) What assumptions are required for this test? Do you feel that these assumptions are reasonable here? Explain briefly.

   Solution:
   We would need the two samples to be independent and each sample to be an independent, random sample from a normal population of run times. Since two separate random samples were obtained 10 years apart, there may not be any issues with independence. It is likely, however, that the run times are right skewed, though this would not be obvious without additional subject knowledge. Specifically, "hour-long" cable television shows

often actually run longer than an hour (on AMC, FX, etc.) with commercials usually making up a similar proportion of the overall length. Thus, while these shows are always at least a certain length (more than 40 minutes not counting commercials), they can sometimes be significantly longer.

Also note that if the 10 most popular "hour-long" cable shows are not representative of all popular "hour-long" cable television shows (in terms of run times), then our inference is somewhat restricted with respect to the critic's original goal. Finally, note that the critic would have to define "popular," which makes the target population for inference unclear as well.

2. (30 points) Wearing a mask helps to control seasonal influenza virus transmission, but each mask is only effective for a certain period of time. Professor X conducted a two-factor balanced experiment using a completely randomized design to study how the duration of time that a specific mask remains effective changes depending on temperature and humidity levels. She studied two levels of temperature: cold (40 degrees F) and warm (80 degrees F). She also studied two levels of humidity: low (60%) and high (85%). Two masks were tested for each combination of temperature and humidity. The measured times (in hours) are given below. Professor X fit an effects model with interaction.

	Humidity	
Temperature	low	high
cold	10, 16	5, 7
warm	6, 11	3, 5

(a) (10 points) Write down the mathematical model for an effects model with interaction, listing all assumptions and explaining all terms.

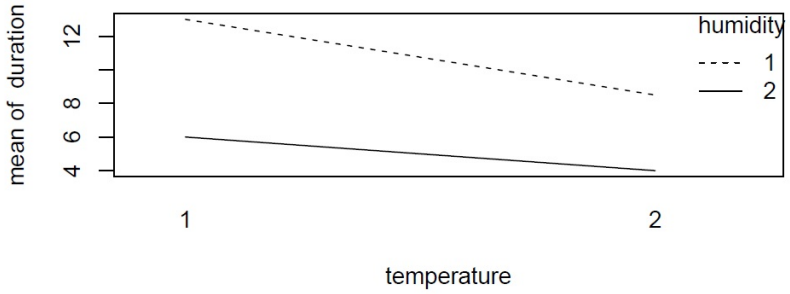
Solution:

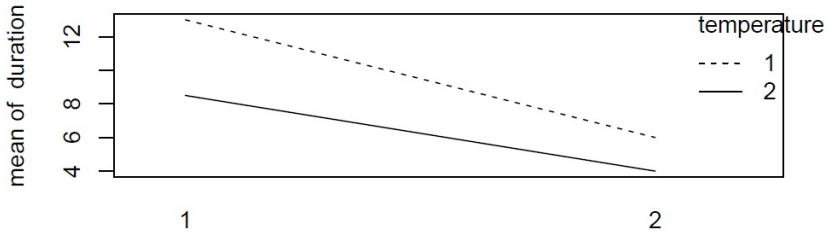
The model is $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$, where the ϵ_{ijk} values are iid $N(0, \sigma^2)$. Here y_{ijk} is the kth observation on level i of temperature and level j of humidity, μ is the grand mean parameter, α_i is the effect of level i of temperature, β_j is the effect of level j of humidity, γ_{ij} is the interaction effect for level i of temperature and level j of humidity, and ϵ_{ijk} is the error term for the kth observation on level i of temperature and level j of humidity. Each of the indices i, j, and k takes on the values 1 and 2.

(b) (10 points) Create interaction plots and comment on what you see.

Solution:

The cell means are 13.0 (cold and low), 6.0 (cold and high), 8.5, (warm and low), and 4.0 (warm and high). Both possible interaction plots are given in the figure. Though the lines are not perfectly parallel, they are close to parallel, with no crossing. Thus, there is no evidence of disorderly interaction and only minimal evidence of any interaction at all between temperature and humidity.





(c) (10 points) Using level 0.05, test for interaction.

Solution:

We need to test $H_0 : \gamma_{11} - \gamma_{12} - \gamma_{21} + \gamma_{22} = 0$ against $H_a : \gamma_{11} - \gamma_{12} - \gamma_{21} + \gamma_{22} \neq 0$. The test statistic is

$$t = \frac{\bar{y}_{11} - \bar{y}_{12} - \bar{y}_{21} + \bar{y}_{22}}{\sqrt{MSE\left(\frac{1^2}{2} + \frac{(-1)^2}{2} + \frac{(-1)^2}{2} + \frac{1^2}{2}\right)}},$$

which under H_0 would follow a t distribution with $8 - 1 - 1 - 1 - 1 = 4$ degrees of freedom. Since the model includes interaction, we can recover SSE by computing the sum of the squared deviations from the cell sample means. This gives

$$\begin{aligned} SSE &= (10-13)^2 + (16-13)^2 + (5-6)^2 + (7-6)^2 + \\ &\quad (6-8.5)^2 + (11-8.5)^2 + (3-4)^2 + (5-4)^2 = 34.5, \end{aligned}$$

meaning that $MSE = SSE/df = 34.5/4 = 8.625$. The test statistic is then

$$t = \frac{13 - 6 - 8.5 + 4}{\sqrt{8.625(2)}} \approx 0.60.$$

Since $|0.60| < t_{0.025,4} \approx 2.776$, we retain H_0. We cannot conclude at level 0.05 that there is an interaction between temperature and humidity.

3. (15 points) For the 2019-2020 year, Villanova published the following enrollment data classifying degree-seeking undergraduate students based on race/ethnicity and class year (rounded to the nearest multiple of 10):

	Race/Ethnicity				
Class Year	White	Hispanic	Asian	Black	Other
First-year	1200	180	120	80	120
Sophomores or higher	3780	380	260	270	390

Conduct a test at the 0.05 level to determine if the distribution of students by race/ethnicity is the same for first-year students and those who are sophomores or higher.

Solution:

H_0 : The distribution of race/ethnicity is the same for both first-year students and sophomores or higher ($\pi_{1,\mathrm{W}} = \pi_{2+,\mathrm{W}}, \pi_{1,\mathrm{H}} = \pi_{2+,\mathrm{H}}, \pi_{1,\mathrm{A}} = \pi_{2+,\mathrm{A}}, \pi_{1,\mathrm{B}} = \pi_{2+,\mathrm{B}}, \pi_{1,\mathrm{O}} = \pi_{2+,\mathrm{O}}$, where $\pi_{1,\mathrm{W}}$ and $\pi_{2+,\mathrm{W}}$ are the proportions first-year students and sophomores or higher who are white, respectively, with the other proportions are defined similarly)

vs.

H_a : The distribution of race/ethnicity is different for first-year students and sophomores or higher (at least one of the equalities in H_0 fails).

Expected values for cells under homogeneity:

	Race/Ethnicity				
Class Year	White	Hispanic	Asian	Black	Other
First-year	1248.7	140.4	95.3	87.8	127.9
Sophomores or higher	3731.3	419.6	284.7	262.2	382.1

All expected values are well greater than 5, and thus our sample size is sufficiently large to conduct Pearson's Chi-square test.

$$\chi^2 - \text{stat} = \frac{(1200 - 1248.7)^2}{1248.7} + \cdots + \frac{(390 - 382.1)^2}{382.1}$$
$$\approx 1.90 + 11.16 + 6.41 + 0.69 + 0.49 + 0.63 + 3.73 + 2.14 + 0.23 + 0.16 = 27.55$$

Critical/Rejection Region:$\chi^2 - \text{stat} > \chi^2_{4,0.05} = 9.49$

Thus, since our test statistic is in the rejection region, we have enough evidence at the 0.05 to conclude that the distribution of race/ethnicity is different for first-years and those in higher class years.

4. (25 points) The Australian fifty-cent coin has the face of Queen Elizabeth II on one side (heads) and the Australian coat of arms on the other side (tails). Suppose that one of these Australian fifty-cent coins was tossed 200 times, producing 84 tails.

   (a) (10 points) Construct a 99% confidence interval for the proportion of times that the Australian fifty-cent coin lands tails up. List any necessary assumptions.

   Solution:
   Assumptions: We assume that the 200 coin tosses are independent with the same probability of landing tails up each toss. We also assume that our sample sizes are large enough such that the proportion of tails up is approximately normal (this should be reasonable; checking the usual condition with the data, $200 * 0.42 = 84$ and $200 * 0.58 = 116$ are both well greater than 5).

   95% CI: $0.42 \pm 2.575\sqrt{\frac{0.42*0.58}{200}} = (0.330, 0.510)$

   Thus, we are 95% confident that the true proportion of times that this coin lands tails up is between 0.330 and 0.510.

   (b) (10 points) Suppose that an 80% confidence interval for the proportion of times that the Australian fifty-cent coin lands tails up is approximately (38%, 46%). Based on this interval, answer each of the following as "True" or "False".

      i. In this experiment, there is an 80% probability that this coin landed tails up between 38% and 46% of the tosses.

      Solution:
      False. We observed 42% tails in THIS experiment. Thus, the probability is 100%, not 80%, that we observed between 38% and 46% tosses resulting in tails up. Even if the question had used the word "confident" instead of "probability", it would not be correct to say that we are confident about the results of a particular experiment (instead, we are confident in the process leading to an interval including the true proportion 80% of the time).

      ii. We can conclude at the 0.2 significance level that this coin does not land heads up 50% of the time.

      Solution:
      True. We can conduct a 0.2 level two-sided hypothesis test of $H_0 : \pi = 0.5$ vs. $H_a : \pi \neq 0.5$ using an 80% CI. Since 0.5 is not in the

confidence interval, we can say that the percent of heads is not 50.

Note: Technically, since the standard error of the CI is based on $\hat{\pi}$, there could be a case where the interval only barely includes (or fails to include) $\pi_0 = 0.5$ and produces a different answer than a hypothesis test using the standard error based on $\pi_0 = 0.5$. However, this is not of concern in this case, since 0.5 is two margins of error above 0.42.

iii. The margin of error of this confidence interval is approximately 4%.

Solution:
True. $(0.46 - 0.38)/2 = 0.04$

iv. If the Australian fifty-cent coin is tossed another 100 times, we are 80% sure that between 38 and 46 of the tosses will be tails.

Solution:
False. Even with the word "confident" rather than "sure", we cannot use a CI based on a particular sample to suggest what will happen in future samples.

v. Doubling the sample size to 400 would reduce the standard error.

Solution:
True. This will usually be true whenever the sample size increases, but it is guaranteed when we go from 200 to 400 since for any $\hat{\pi}$, $\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{400}} < \sqrt{\frac{0.42*0.58}{200}}$.

(c) (5 points) Suppose that the p-value for a test of $H_0 : \pi = 0.5$ vs. $H_0 : \pi \neq 0.5$ is 0.024, indicating that we can conclude that the coin is not fair at the 0.05 level. Carefully interpret the meaning of the p-value, 0.024.

Solution:

If the true proportion of times that the coin lands tails up is 0.5, then we would observe a z-stat at least as large in absolute value as we did with probability 0.024. Equivalently, if the true proportion of times that the coin lands tails up is 0.5, then we would observe a a $\hat{\pi}$ at least as far from 0.5 as we did with probability 0.024.

5. (30 points) An ice cream shop owner ran an experiment to study how the melting time for a bowl of ice cream varies depending on the flavor of ice cream and the type of bowl that is used. The owner selected four flavors at random from the many flavors offered by the shop, and he used each of the three types of bowls that the shop owns. For each combination of a flavor and a bowl type, he filled two bowls and recorded the melting time in seconds.

Source	df	SS	MS	Expected MS
Flavor		600		$\sigma^2 + 2\sigma^2_{FB} + 6\sigma^2_F$
Bowl type		200		$\sigma^2 + 2\sigma^2_{FB} + Q(B)$
Interaction		120		$\sigma^2 + 2\sigma^2_{FB}$
Error		60		σ^2

(a) (10 points) Write down an appropriate mathematical model that allows for interaction, listing all assumptions and explaining all terms.

Solution:

The model is $y_{ijk} = \mu + a_i + \beta_j + (a\beta)_{ij} + \epsilon_{ijk}$, where the ϵ_{ijk} values are iid $N(0, \sigma^2)$, the a_i values are iid $N(0, \sigma^2_F)$, the $(a\beta)_{ij}$ values are iid $N(0, \sigma^2_{FB})$, and all of the random variables are mutually independent. Here y_{ijk} is the kth observation on level i of flavor and level j of bowl type, μ is the grand mean parameter, a_i is the random effect of level i of flavor, β_j is the effect of level j of bowl type, $(a\beta)_{ij}$ is the random interaction effect for level i of flavor and level j of bowl type, and ϵ_{ijk} is the error term for the kth observation on level i of flavor and level j of bowl type. The bounds on the indices i, j, and k are 1 to 4 (i), 1 to 3 (j), and 1 to 2 (k).

(b) (5 points) Complete the partial ANOVA table given above. Note that $Q(B)$ is a quadratic form in the bowl effects and that σ^2, σ^2_F, and σ^2_{FB} are variance components for error, flavor, and the interaction, respectively.

Solution:

We get the df for flavor as $4 - 1 = 3$, the df for bowl type as $3 - 1 = 2$, and the df for interaction as $3(2) = 6$. The error df can be obtained by subtraction as $4(3)(2) - 1 - 3 - 2 - 6 = 12$ or from noting that there is one error df for each choice of a flavor and a bowl type. This gives the table below.

Source	df	SS	MS	Expected MS
Flavor	3	600	200	$\sigma^2 + 2\sigma^2_{FB} + 6\sigma^2_F$
Bowl type	2	200	100	$\sigma^2 + 2\sigma^2_{FB} + Q(B)$
Interaction	6	120	20	$\sigma^2 + 2\sigma^2_{FB}$
Error	12	60	5	σ^2

(c) (15 points) Make appropriate inferences. Specifically, test the significance of all fixed effects using level 0.05, and estimate all variance components.

Solution:

By the method of moments, our estimate for the error variance is $\hat{\sigma}^2 = 5$. Bringing in the interaction line then gives us that $\hat{\sigma}^2 + 2\hat{\sigma}^2_{FB} = 20$, which implies that $\hat{\sigma}^2_{FB} = 7.5$. Bringing in the flavor line then gives that $\hat{\sigma}^2_F = (200 - 20)/6 = 30$.

We can test $H_0 : \beta_1 = \beta_2 = \beta_3$ against H_a : The β_i values are not all equal by using $F = MSB/MSFB$, which is appropriate since the expected MS values are the same except for $Q(B)$, which is zero iff H_0 holds. This gives $F = 100/20 = 5$, which we compare to $F_{0.05,2,6} \approx 5.14$. Since $5 < 5.14$, we do not have enough evidence to conclude at level 0.05 that there is a difference between the bowl types in terms of mean melting time.

6. (30 points) A Karnofsky Score ranges from 0 to 100 and is meant to rate a patient's ability to carry out daily activities (with 100 indicating a high ability). In a particular study on 100 lung cancer patients, suppose that each patient and their doctor were asked to assign a Karnofsky score for the patient. The patients assigned themselves an average Karnofsky score of 83 with a standard deviation of 8 while doctors assigned the patients an average Karnofsky score of 84 with a standard deviation of 7. The differences in the 100 scores (patient minus doctor) have a mean of -1 and a standard deviation of 5.

   (a) (10 points) Can we conclude at the 0.05 level that lung cancer patients assign themselves lower Karnofsky scores than their doctors on average? Conduct the appropriate hypothesis test.

   Solution:
   $H_0 : \mu_d \geq 0$ vs. $H_a : \mu_d < 0$, where μ_d is the mean in the differences of the patient and doctor (patient minus doctor) Karnofsky scores.

   t-stat $= \frac{-1-0}{5/\sqrt{100}} = -2.00$

   Critical/Rejection Region: t-stat $< -t_{99,0.05} = -1.66 \approx -t_{60,0.05} = -1.671$.

   Thus, since our test statistic is in the rejection region, we have enough evidence at the 0.05 level to conclude that patients assign themselves lower Karnofsky scores than their doctor, on average.

   (b) (5 points) What assumptions are needed for the test in part (a)? Do you feel that these assumptions are reasonable here? Explain briefly.

   Solution:
   We need to assume that we have an independent, random sample of differences in Karnofsky scores. We also need for the differences in the Karnofsky scores to be approximately normal, or, alternatively, that our sample size is large enough to overcome any significant skew. While not enough details are given to answer whether the sample is approximately independent (or at least representative), there could be many reasons the sample would not be good: groups of patients from the same hospital/clinic, patients sharing the same doctor, certain patients receiving differing levels of instruction before assigning a score, etc. Since our sample size of 100 is relatively large, as long as the differences in the scores aren't very skewed, the t-statistic should be approximately t_{99} under H_0 assuming no sampling issues.

   (c) (5 points) What sample size would be needed if we wanted 80% power for the test in part (a) if the true difference in mean Karnofsky scores (patient

minus doctor) is -0.5 and the true standard deviation of the differences is 6.

Solution:
We can use the standard formula: $n = \frac{\sigma_d^2(z_\beta + z_\alpha)^2}{\delta^2}$, where δ is difference in the mean scores for which we desire a particular power. In this case, this gives $n = \frac{6^2(0.84+1.645)^2}{(-0.5)^2} = 889.23$. Thus, we would need 890 lung cancer patients to achieve the desired power for this test. We can also derive the sample size as follows:

$$\begin{aligned}
P(\text{Z-stat} < -1.645 | \mu_d = -0.5) = 0.8 &\implies \\
P\left(\frac{\overline{Y} - 0}{6/\sqrt{n}} < -1.645 | \mu_d = -0.5\right) = 0.8 &\implies \\
P\left(\frac{\overline{Y} - (-0.5)}{6/\sqrt{n}} < -1.645 + \frac{0.5}{6/\sqrt{n}} | \mu_d = -0.5\right) = 0.8 &\implies \\
P(Z < -1.645 + 0.5\sqrt{n}/6) = 0.8 &\implies \\
-1.645 + 0.5\sqrt{n}/6 = 0.84 &\implies \\
n = \frac{(0.84 + 1.645)^2 6^2}{(0.5)^2} = 889.23 &
\end{aligned}$$

(d) (10 points) Suppose that the American Medical Association considers Karnofsky scores for a population of patients unreliable if for a sample of $n \geq 100$ the average sample difference in Karnofsky scores between patient and doctor is at least 2 in absolute value. What is the chance that Karnofsky scores are found to be unreliable based on a sample of 100 lung cancer patients if the true average difference in scores is 1 and the true standard deviation of the differences is 6?

Solution:

$$\begin{aligned}
Power &= P(|\overline{Y}| > 2 | \mu_d = 1) \\
&= P(\overline{Y} > 2 | \mu_d = 1) + P(\overline{Y} < -2 | \mu_d = 1) \\
&= P\left(Z > \frac{2 - (-1)}{6/\sqrt{100}}\right) + P\left(Z < \frac{-2 - (-1)}{6/\sqrt{100}}\right) \\
&= P(Z > 5) + P(Z < -1.\overline{6}) \approx 0 + 0.048 = 0.048
\end{aligned}$$

7. (55 points) The World Happiness Report is conducted each year. It compares the level of happiness of people from all the countries of the world, along with other variables. The following analyses were done using a subset of the data in a given year where the case is a country (there is one observation per country). Specifically, they considered the following variables in a linear regression equation to estimate happiness.

Variable	Description
happy	Average happiness score among all respondents in that country. Each person was told "Imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?"
logGDP	Natural log of gross domestic product per capita (in 2011 US$)
socsup	Proportion of respondents in that country who responded "yes" to the question "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"
sshigh	"TRUE" if socsup for the country was above the median value of socsup "FALSE" if socsup for the country was at or below the median value of socsup

Using the output, answer the following questions:

(a) (5 points) How many countries were included in Model 1? Show your work.

Solution:
$df_E = 116$. Since there are two parameters in the model (Slope and Intercept), this equals $n - 2$, so $n = 118$.

(b) (5 points) Why is the slope for *logGDP* for Model 1 different than the slope for *logGDP* for Model 2?

Solution:
The slope for *logGDP* in Model 2 controls or adjusts for *sshighTRUE*, while the slope for *logGDP* in Model 1 does not. When you add or remove a variable from the model, the slopes and their interpretations change (assuming they are not orthogonal).

(c) Regarding the values of R^2:

i. (5 points) Identify and interpret the value of R^2 in Model 1.

Solution:

$R^2 = 0.6019$. This means that approximately 60% of the variation in happiness can be explained by the linear regression on *logGDP*.

ii. (5 points) Using information from Model 1, what are the possible value of the R^2 value for Model 2? Explain your answer.

Solution:
Since you added a variable to Model 2, the R^2 must be at least as high as in Model 1. So the possible values are $[0.6019, 1]$.

(d) (15 points) For Model 2, conduct the appropriate hypothesis test for *sshighTRUE*. Explain why this test is of interest.

Solution:
$H_0 : \beta_{sshighTRUE} = 0$ vs. $H_a : \beta_{sshighTRUE} \neq 0$

$t_{115} = 0.7074/0.1661 = 4.259$

$P(|t_{115}| > 4.259) = 4.2 \times 10^{-5}$, or 0.000042. (Alternatively, $4.259 > 1.98 = t_{115,0.975}$). Since the p-value is less than 0.05 and the test statistic is in the rejection region, we reject the null in favor of the alternative. That is, at the 5% level of significance, there is enough evidence to conclude that the slope for *sshigh* is non-zero, when controlling for *logGDP*.

This test is of interest if we want to know whether there is a linear relationship between *sshigh* and *happiness*, when controlling for *logGDP*. Since *sshigh* is an indicator variable, we are really determining whether the expected happiness score is the same for *sshighTRUE* and *sshighFALSE*, when controlling for *logGDP*. If it were not significant, we might consider dropping *sshigh* out of the model.

(e) (15 points) There are two graphs for Model 2. What assumptions can be explored with these graphs? For each assumption, comment on whether it is satisfied based on these graphs.

Solution:

Graph	Assumption	Comment
Normal Q-Q	Normal Errors	The points look to be fairly close to the line, so the normality of errors assumptions appears to be reasonable.
Residual	Constant Variance	The points seem mostly randomly scattered around the line – there is no funnel or barbell shape. This assumption appears to be reasonable.
Residual	Linearity	The points are mostly scattered evenly above and below the line throughout, with the possible exception of no points below the line at the end (consistent with the higher red (LOESS) line). There is slight evidence of non-linearity towards the high predicted values.
Residual	No Influential Points? (technically not an assumption, but an important condition for the model to be reasonable/stable)	The residual plot can sometimes be used to suggest the influence of some points. In this plot, none stand out.

(f) (5 points) In Model 3, the line *logGDP::sshighTRUE* is the interaction effect between *logGDP* and *sshighTRUE*. Interpret its estimate of 0.42030.

Solution:

When predicting happiness using this model, the slope for logGDP is 0.42030 higher for sshighTRUE than it is for sshighFALSE.

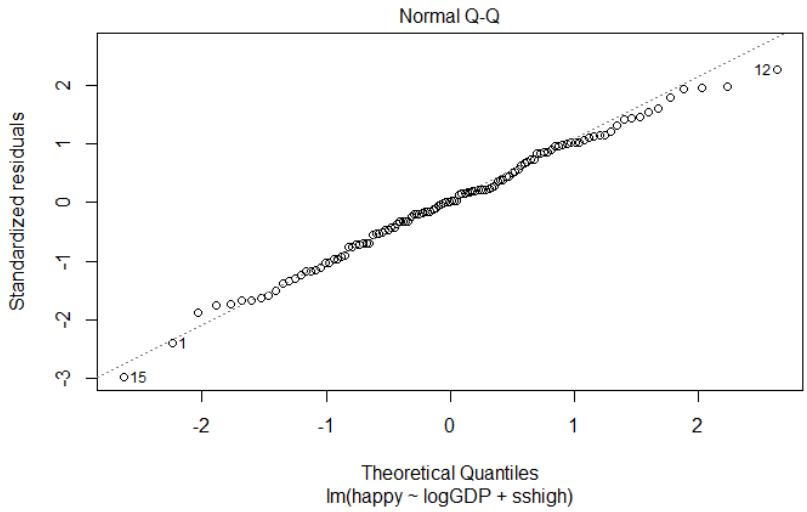
Happiness - Model 1

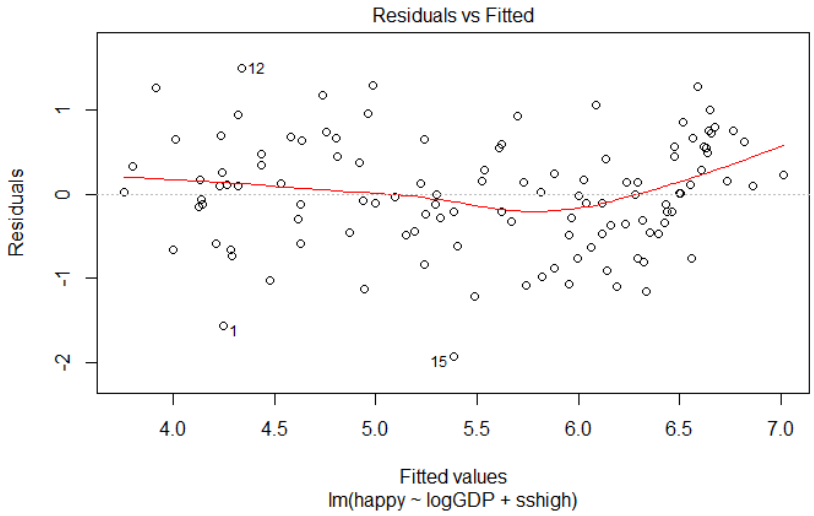
```
Coefficients:
             Estimate Std. Error
(Intercept) -1.15776    0.50853
logGDP       0.72273    0.05458

Residual standard error: 0.704 on 116 degrees of freedom
Multiple R-squared:  0.6019,    Adjusted R-squared:  0.5984
F-statistic: 175.4 on 1 and 116 DF,  p-value: < 2.2e-16
```

Happiness - Model 2

```
Coefficients:
             Estimate Std. Error
(Intercept)   0.3605     0.5937
logGDP        0.5189     0.0699
sshighTRUE    0.7074     0.1661
```





Happiness - Model 3

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.56247    0.70520   2.216  0.02870 *
logGDP              0.37580    0.08334   4.509 1.59e-05 ***
sshighTRUE         -3.27449    1.36280  -2.403  0.01789 *
logGDP:sshighTRUE   0.42030    0.14284   2.942  0.00395 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```