STAT021 Statistical Methods II

Lecture 14 Multiple Linear Regression

Lu Chen
Swarthmore College
10/30/2018

Multiple Linear Regression

CHOOSE

- Exploratory data analysis
- Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \epsilon$, where $\epsilon \overset{iid}{\sim} N(0, \sigma)$

FIT

- Maximum likelihood estimation (MLE)

ASSESS model

- Inference for the intercept and slopes; ANOVA and R^2

ASSESS error

- Check conditions; Unusual points

USE

- Predictions

Multiple Linear Regression

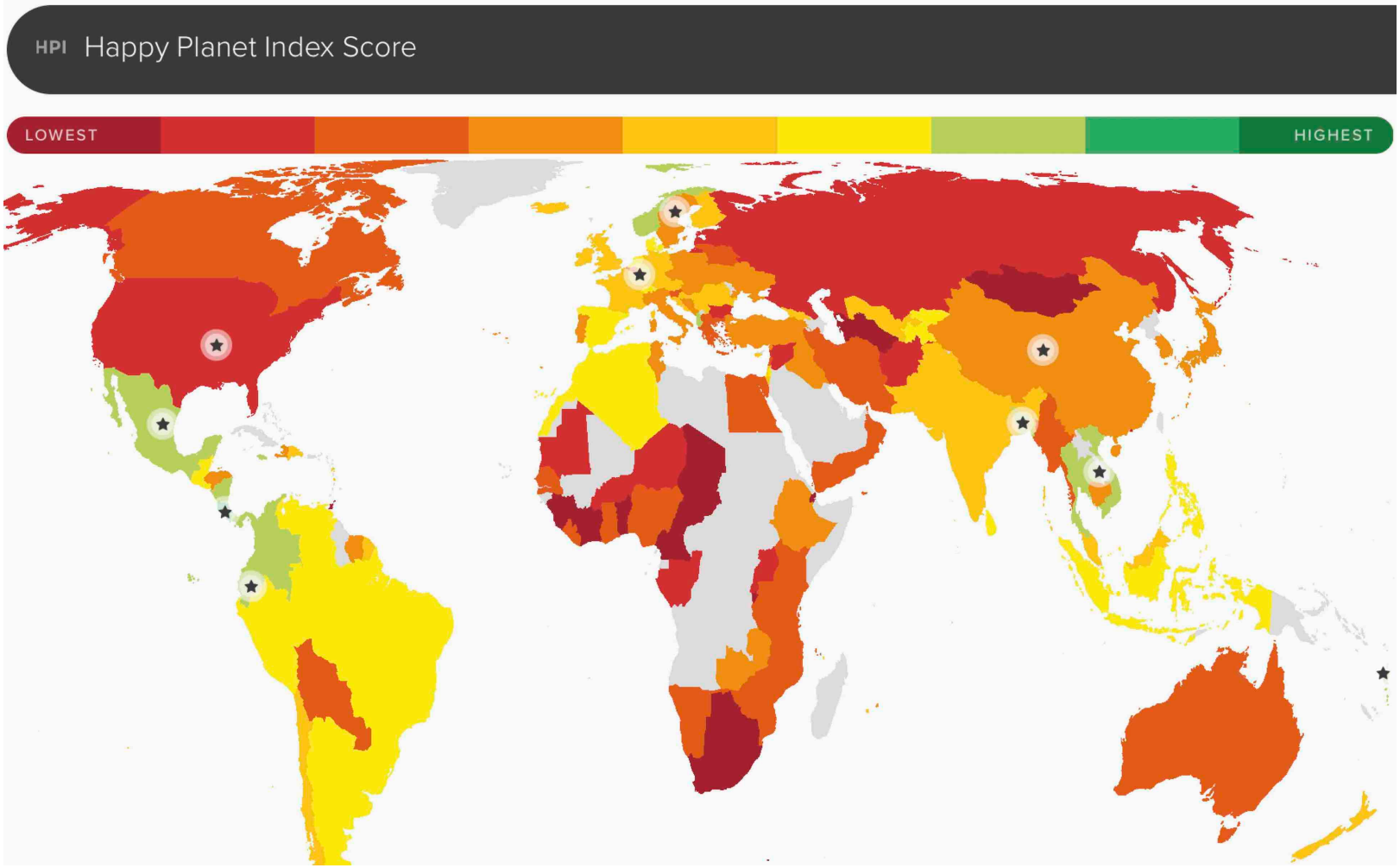
New topics

- Nested F test for a subset of predictors
- Adjusted R^2 for model comparions
- Categorical predictors
- Interaction between predictors
- Transformation and polynomial regression
- Multicollinearity
- Model building
  - Forward selection
  - Backward elimination
  - Stepwise procedure

Outline

- **CHOOSE**
  - Exploratory data analysis
  - Model definition
- **FIT**: Maximum likelihood estimation (MLE)
- **ASSESS model**
  - Inference for the intercept and slopes
- **ASSESS error**: Check conditions; unusual points
- **USE**: Predictions

Happy Planet Index



Link

CHOOSE: Exploratory data analysis

1. Summary statistics of indivdual variables

```
head(HappyPlanet, 3)
```

```
##       Country Happiness      GDPpc LifeExp
## 1 Philippines  59.17430  1678.8520    70.4
## 2      Rwanda  28.34747   398.2085    43.9
## 3     Hungary  37.63759 13842.6055    72.7
```

```
dim(HappyPlanet)
```

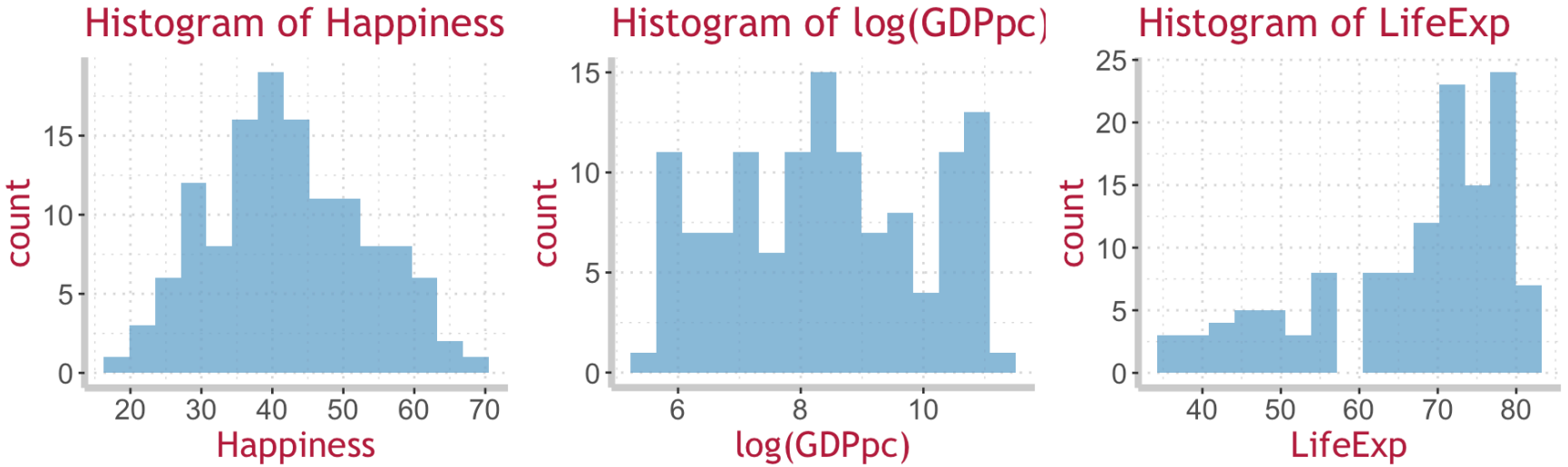
```
## [1] 128   4
```

```
library(psych) # package for the describe() function
describe(HappyPlanet)[,2:4] # summary statistics
```

```
##                n     mean       sd
## Country*     128    68.13    39.85
## Happiness    128    42.29    10.99
## GDPpc        124 13066.54 17971.54
## LifeExp      128    66.72    12.17
```

CHOOSE: Exploratory data analysis

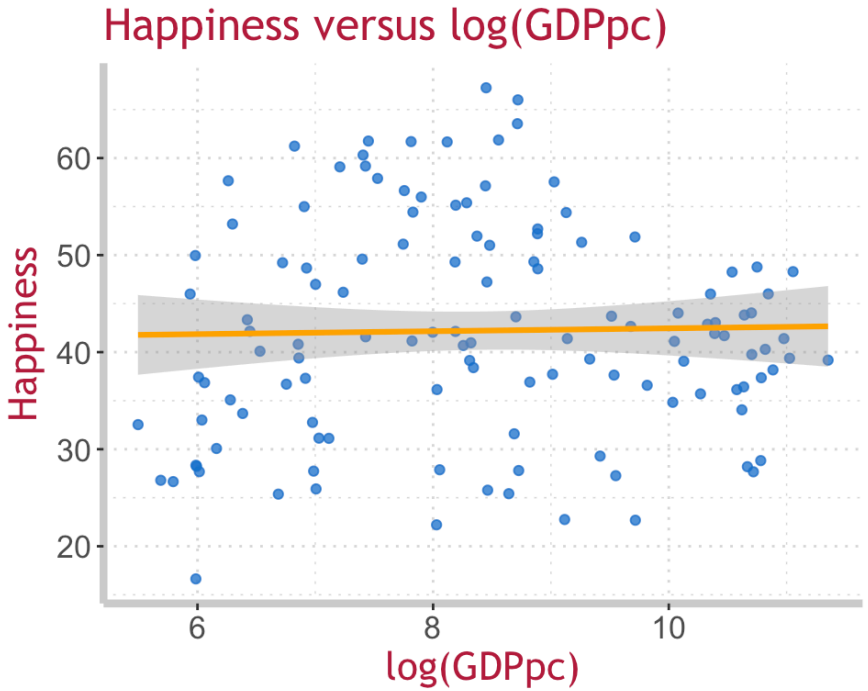
2. Distributions of indivdual variables

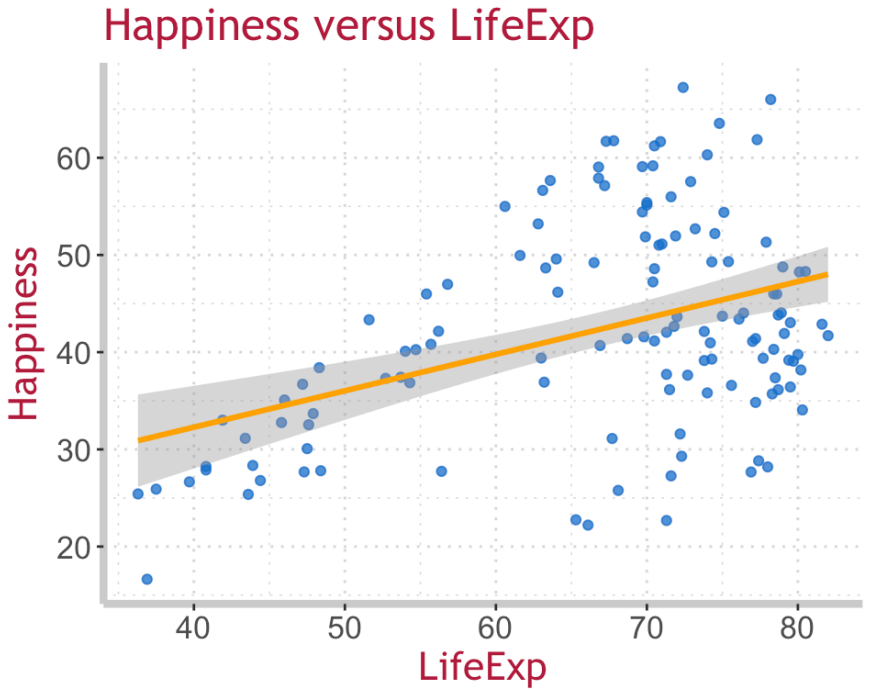


- *Happiness*: response variable, close to Normal.
- *log(GDPpc)*: explanatory variable, quite uniform.
- *LifeExp*: explanatory variable, left skewed.

CHOOSE: Exploratory data analysis

3. Scatterplots of the response variables versus each explanatory variable





- *Happiness* versus *log(GDPpc)*: positive, linear, extremely weak
- *Happiness* versus *LifeExp*: positive, linear, moderate

CHOOSE: Exploratory data analysis

4. SLR of the response variables versus each explanatory variable

```
m1 <- lm(Happiness ~ log(GDPpc), data=HappyPlanet)
m2 <- lm(Happiness ~ LifeExp, data=HappyPlanet)
summary(m1)
summary(m2)
```

Models	Estimated regression line	t test for slope	R^2
1. *Happiness ~ log(GDPpc)*	$\hat{y} = 40.95 + 0.15x$	$t = 0.24$, $P = 0.811$	0.00047
2. *Happiness ~ LifeExp*	$\hat{y} = 17.30 + 0.37x$	$t = 5.12$, $P = 1.14 \times 10^{-6}$	0.17

- The linear relationship between *Happiness* and *log(GDPpc)* is not significant.
- *Happiness* and *LifeExp* have highly significant linear relationship.

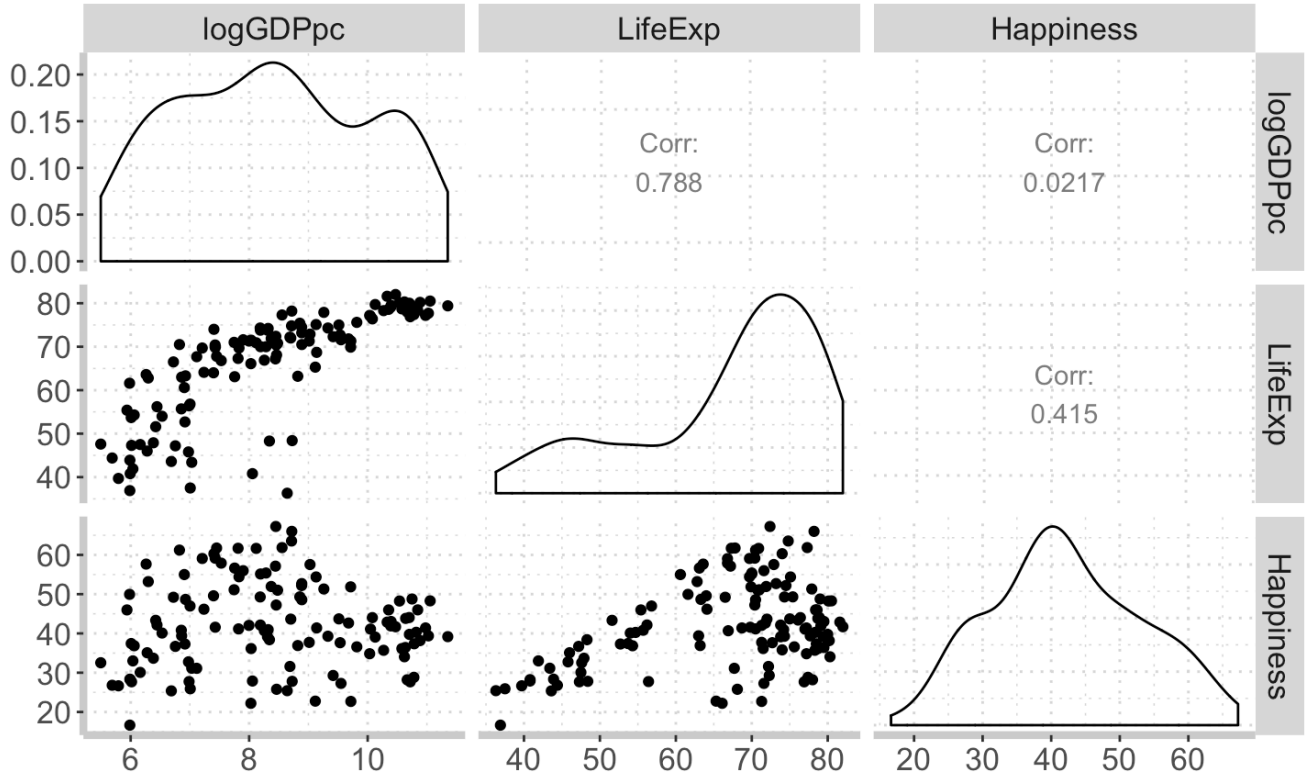
CHOOSE: Exploratory data analysis

1. Summary statistics of indivdual variables
2. Distributions of indivdual variables
3. Scatterplots of the response variables versus each explanatory variable
4. SLR of the response variables versus each explanatory variable

- Note, at this point, usually we do not check model assumptions for the SLR models because our interest is an MLR model including all the explanatory variables. We will check the model assumptions for the MLR model later.

CHOOSE: Exploratory data analysis

```
HappyPlanet <- data.frame(HappyPlanet, logGDPpc=log(HappyPlanet$GDPpc))
library(GGally)
ggpairs(data=HappyPlanet[, c("logGDPpc", "LifeExp", "Happiness")])
```



Instead of generating individual histograms and scatterplots, you may use the `ggpairs()` function from the `GGally` package to visualize the data easily.

CHOOSE: Model definition

For n observations on K explanatory variables $X_1, X_2, \cdots, X_K$ and a response variable Y, to study or predict the behavior of Y for the given set of the explanatory variables, the **multiple linear regression model** is defined as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \epsilon$$

where $\epsilon \overset{iid}{\sim} N(0, \sigma)$.

- $\mu_Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$
  Mean of Y is a function of the explanatory variables $X_1, X_2, \cdots, X_K$
- $Y = \mu_Y + \epsilon$ and $\epsilon \overset{iid}{\sim} N(0, \sigma)$
  SD of ϵ is a constant.
- $K + 2$ parameters: $\beta_0, \beta_1, \beta_2, \cdots, \beta_K, \sigma$. For SLR, $K = 1$ and it has 3 parameters.
- SLR is a special case of MLR.

CHOOSE and FIT

For our example, denote *Happiness* as Y, $log(GDPpc)$ as X_1 and *LifeExp* as X_2,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

where $\epsilon \overset{iid}{\sim} N(0, \sigma)$.

- 4 parameters: $\beta_0, \beta_1, \beta_2$ and σ.
- These parameters are estimated using **maximum likelihood** method by searching for the values of $\beta_0, \beta_1, \beta_2$ and σ that maximize the likelihood of observing Y given the parameters.
- Estimates: b_0, b_1, b_2 and $\hat{\sigma}$. Estimated regression line $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$ and

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-K-1}} = \sqrt{\frac{\sum (y-\hat{y})^2}{n-K-1}} = \sqrt{\frac{\sum (y - b_0 - b_1 x_1 - b_2 x_2)^2}{n-2-1}}$$

  The degree of freedom of an MLR model is $n - K - 1$.

FIT

```
m3 <- lm(Happiness ~ log(GDPpc) + LifeExp, data=HappyPlanet)
summary(m3)
```

```
## Call: lm(formula = Happiness ~ log(GDPpc) + LifeExp, data = HappyPlanet)
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.64252    4.33521   5.915 3.16e-08 ***
## log(GDPpc)  -5.68509    0.76664  -7.416 1.82e-11 ***
## LifeExp      0.96482    0.09991   9.657  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.371 on 121 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.4355, Adjusted R-squared:  0.4262
## F-statistic: 46.68 on 2 and 121 DF,  p-value: 9.436e-16
```

- $b_0 = 25.6$
- $b_1 = -5.7$
- $b_2 = 1.0$
- $\hat{\sigma} = 8.4$ with $n - K - 1 = 124 - 2 - 1 = 121$ degrees of freedom.

- **Estimated regression line**

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 = 25.6 - 5.7 x_1 + 1.0 x_2$$

ASSESS Model: Inference for intercept & slopes

Individual t tests for the slopes in MLR

To test the coefficient for one of the predictors, X_k, in a multiple linear regression model, the hypotheses are $H_0 : \beta_k = 0$ vs $H_a : \beta_k \neq 0$ and the test statistic is

$$t = \frac{b_k}{SE_{b_k}} \sim t(n - K - 1)$$

Level C confidence intervals for the intercept and slopes are

$$b_0 \pm t^* SE_{b_0} \text{ and } b_k \pm t^* SE_{b_k}$$

$k = 1, 2, \cdots, K$. If the conditions for the MLR model hold, we compute the P-value and t^* value using a t-distribution with $n - K - 1$ degrees of freedom.

ASSESS Model: Slopes and *t* tests

```
summary(m3)$coefficient
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.64252    4.33521   5.915 3.16e-08 ***
## log(GDPpc)  -5.68509    0.76664  -7.416 1.82e-11 ***
## LifeExp      0.96482    0.09991   9.657  < 2e-16 ***
```

$$\widehat{Happiness} = 25.6 - 5.7 \times log(GDPpc) + 1.0 \times LifeExp$$

- **Slope of *log(GDP)***: $b_1 = -5.7, SE_{b_1} = 0.8$

  Happiness score will be decreased by 5.7 units as *log(GDPpc)* increased by 1 unit, given that *LifeExp* is held constant.

- ***t* test for the slope of *log(GDP)***: $t = b_1/SE_{b_1} = -7.4$ and

  $P = 1.8 \times 10^{-11} < 0.05$

  Given that *LifeExp* is held constant, *log(GDPpc)* has a highly significant negative linear relationship with *Happiness*.

- The *t* test indicates how significant *log(GDPpc)* is when *LifeExp* is considered in the relationship.

ASSESS Model: Slopes and *t* tests

```
summary(m3)$coefficient
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.64252    4.33521   5.915 3.16e-08 ***
## log(GDPpc)  -5.68509    0.76664  -7.416 1.82e-11 ***
## LifeExp      0.96482    0.09991   9.657  < 2e-16 ***
```

$$\widehat{Happiness} = 25.6 - 5.7 \times log(GDPpc) + 1.0 \times LifeExp$$

- **Slope of *LifeExp***: $b_2 = 1.0, SE_{b_2} = 0.1$

  Happiness score will be increased by 1 unit as *LifeExp* increased by 1 unit, given that *log(GDPpc)* is held constant.

- ***t* test for the slope of *LifeExp***: $t = b_2/SE_{b_2} = 10.0$ and $P = 1.1 \times 10^{-16} < 0.05$

  Given that *log(GDPpc)* is held constant, *LifeExp* has a highly significant positive linear relationship with *Happiness*.

- The *t* test indicates how significant *LifeExp* is when *log(GDPpc)* is considered in the relationship.

ASSESS Model: Confidence intervals

```
summary(m3)$coefficient
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.64252    4.33521   5.915 3.16e-08 ***
## log(GDPpc)  -5.68509    0.76664  -7.416 1.82e-11 ***
## LifeExp      0.96482    0.09991   9.657  < 2e-16 ***
```

```
confint(m3)
```

```
##                  2.5 %    97.5 %
## (Intercept) 17.0598368 34.225205
## log(GDPpc)  -7.2028678 -4.167322
## LifeExp      0.7670216  1.162620
```

- The intercept β_0 is estimated as 25.6 with 95% confidence interval [17.1, 34.2]
- The slope for *log(GDP)*, β_1, is estimated as -5.7 with 95% CI [−7.2, −4.2]
- The slope for *LifeExp*, β_2, is estimated as 1.0 with 95% CI [0.8, 1.2]

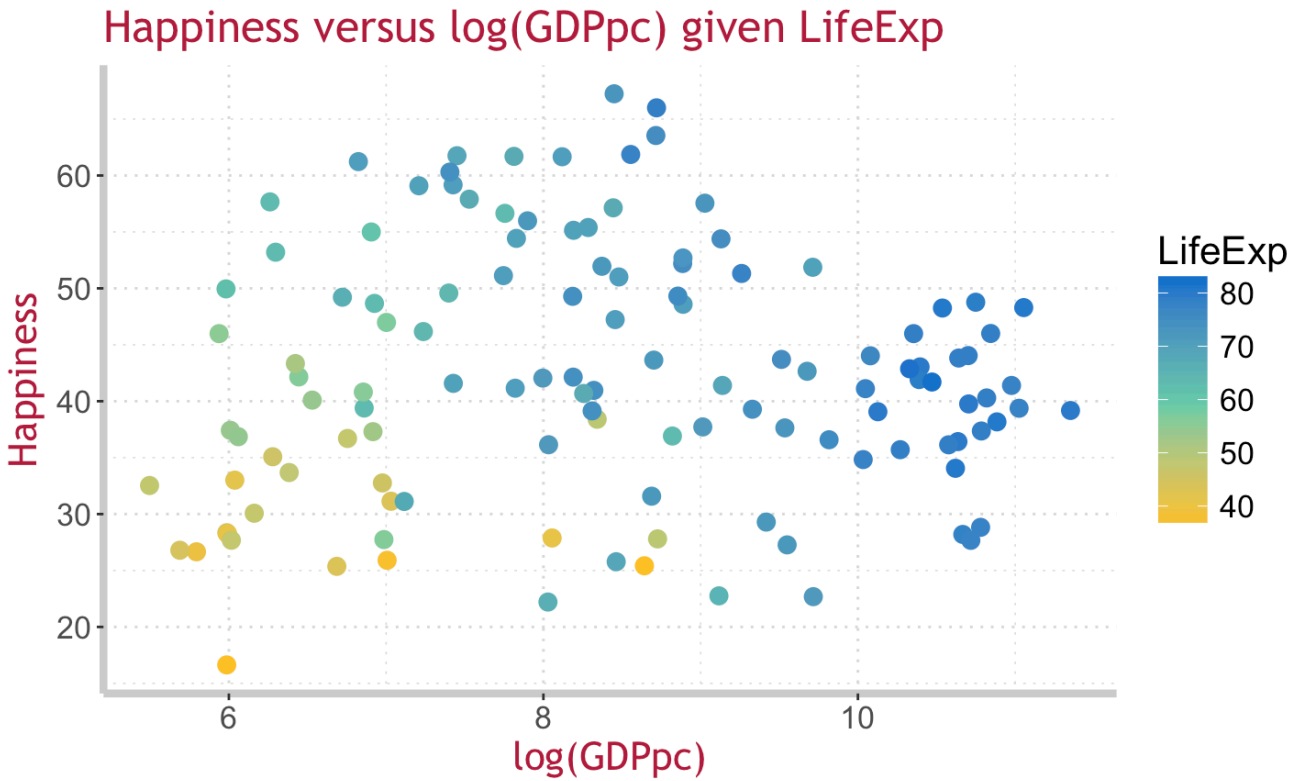
ASSESS model

Models	Estimated regression line	t test for slope	R^2
1. ***Happiness ~ log(GDPpc)***	$\hat{y} = 40.95 + 0.15x$	$P = 0.811$	0.00047
2. ***Happiness ~ LifeExp***	$\hat{y} = 17.30 + 0.37x$	$P = 1.14 \times 10^{-6}$	0.17
3. ***Happiness ~ log(GDPpc) + LifeExp***	$\hat{y} = 25.7 - 5.7x_1 + 1.0x_2$	Both Ps are tiny	0.44

- In model 1, relationship between *Happiness* and *log(GDPpc)* is positive (slope 0.15) and not significant, which becomes negative (slope -5.7) and highly significant in model 3.
- In model 2, the slope for *LifeExp* is 0.37, which becomes much larger (1.0) in model 3.
- Model 1 and 2 have total R^2 17%, while model 3 has R^2 44%.
- How to expain?

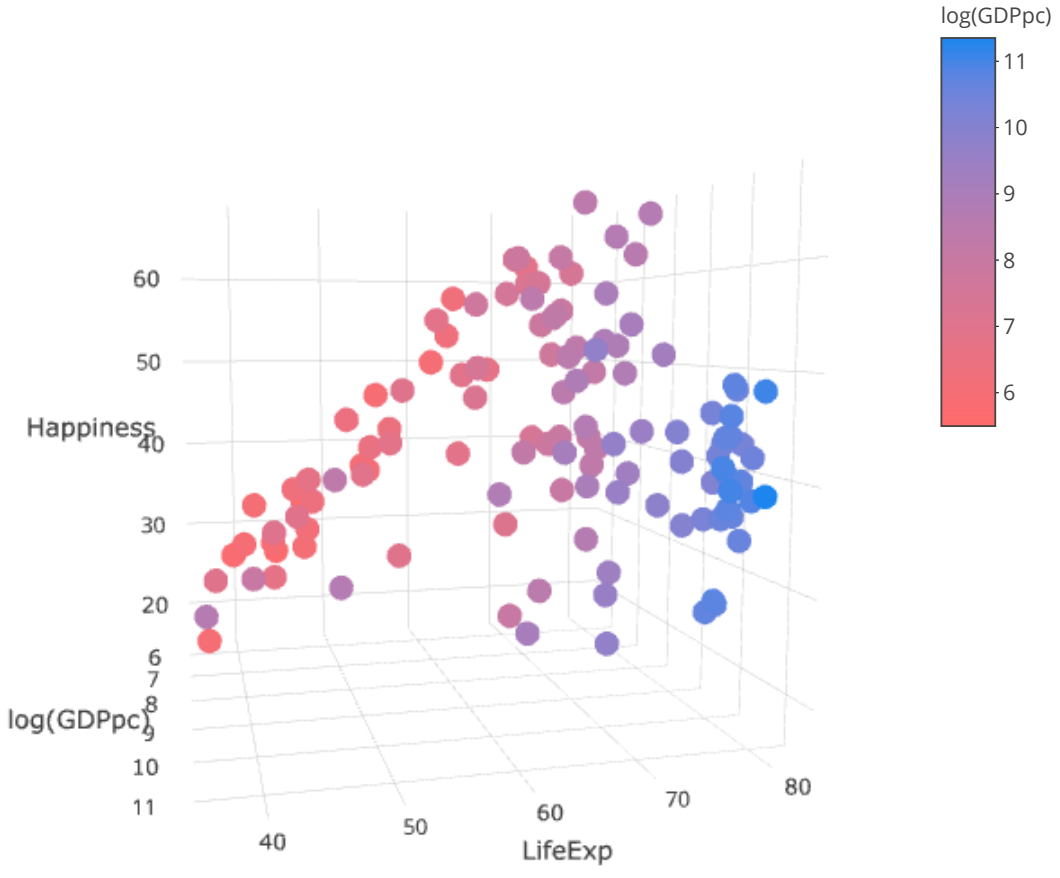
Happiness ~ log(GDPpc) + LifeExp

```
ggplot(HappyPlanet, aes(x=log(GDPpc), y=Happiness, colour=LifeExp))+
  geom_point(size=3)+
  scale_colour_gradientn(colours = c("goldenrod1","aquamarine3","dodgerblue3"))+
  ggtitle("Happiness versus log(GDPpc) given LifeExp")
```



- Countries with higher *GDPpc* also have higher *LifeExp*.
- For countries with similar *LifeExp*, *log(GDPpc)* is negatively related to *Happiness*.

Happiness ~ log(GDPpc) + LifeExp



- For countries with similar *log(GDPpc)* values, *LifeExp* is positively related to *Happiness*. The strength of the relationship after adjusting for *log(GDPpc)* is greater than the overall relationship between *LifeExp* and *Happiness*.

ASSESS error: model assumptions

Linearity: linear relationship between Y and Xs.

- Scatterplot of residuals e on fitted values $\hat{y}$ (pattern?)

Zero mean: mean of the errors is 0 - always true

Constant variance: variability of the errors is the same for all X values.

- Scatterplot of residuals e on fitted values $\hat{y}$ (spread?)
- Breusch-Pagan test for H_0: constant variance

Normality: distribution of the errors is Normal.

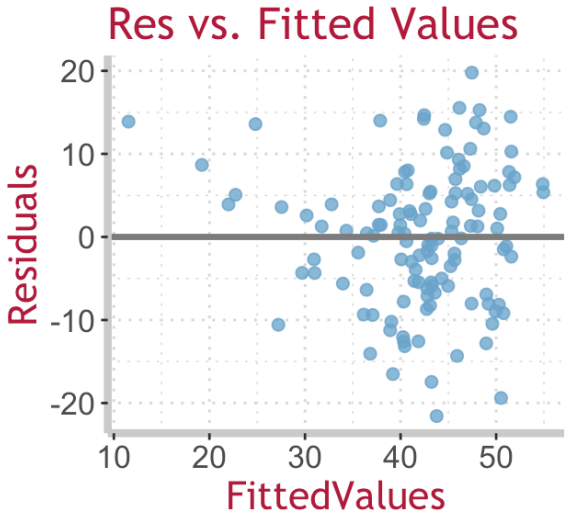
- Normal Q-Q plot (sometimes histogram of residuals is helpful)

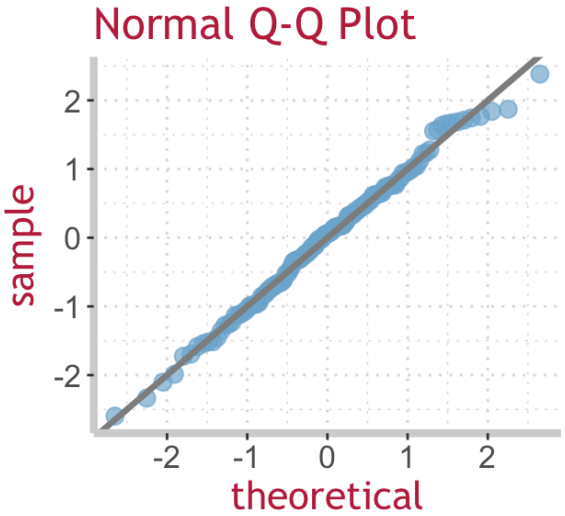
Independence and randomness - check data collecting process

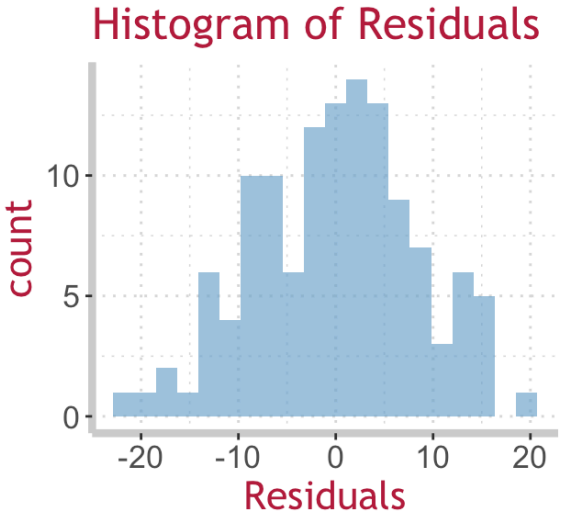
ASSESS error

```
m3 <- lm(Happiness ~ log(GDPpc) + LifeExp, data=HappyPlanet)
Assess <- data.frame(Residuals=m3$residuals,
                     FittedValues=m3$fitted.values)

# Residuals vs. Fitted Values
ggplot(data=Assess, aes(x=FittedValues, y=Residuals))+
  geom_point(color="skyblue3", size=2, alpha=0.8)+
  geom_hline(yintercept=0, size=1.2, colour="grey50")+
  ggtitle("Residuals vs. Fitted Values")

# Normal Q-Q plot
ggplot(data=Assess, aes(sample = scale(Residuals)))+
  stat_qq(size=3, color="skyblue3", alpha=0.7)+
  geom_abline(intercept=0, slope=1, size=1.2, colour="grey50")+
  ggtitle("Normal Q-Q Plot")

# Histogram of residuals
ggplot(Assess, aes(Residuals))+
  geom_histogram(bins=20, fill="skyblue3", alpha=0.7)+
  ggtitle("Histogram of Residuals")
```

ASSESS error: *Happiness~log(GDPpc)+LifeExp*








```
library(lmtest)
bptest(m3) # model Happiness ~ log(GDPpc) + LifeExp
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m3
## BP = 0.45317, df = 2, p-value = 0.7972
```

- No clear violation in model assumptions is found in this MLR model exept for several suspious points in the residuals versus fitted values plot.

Unusual points

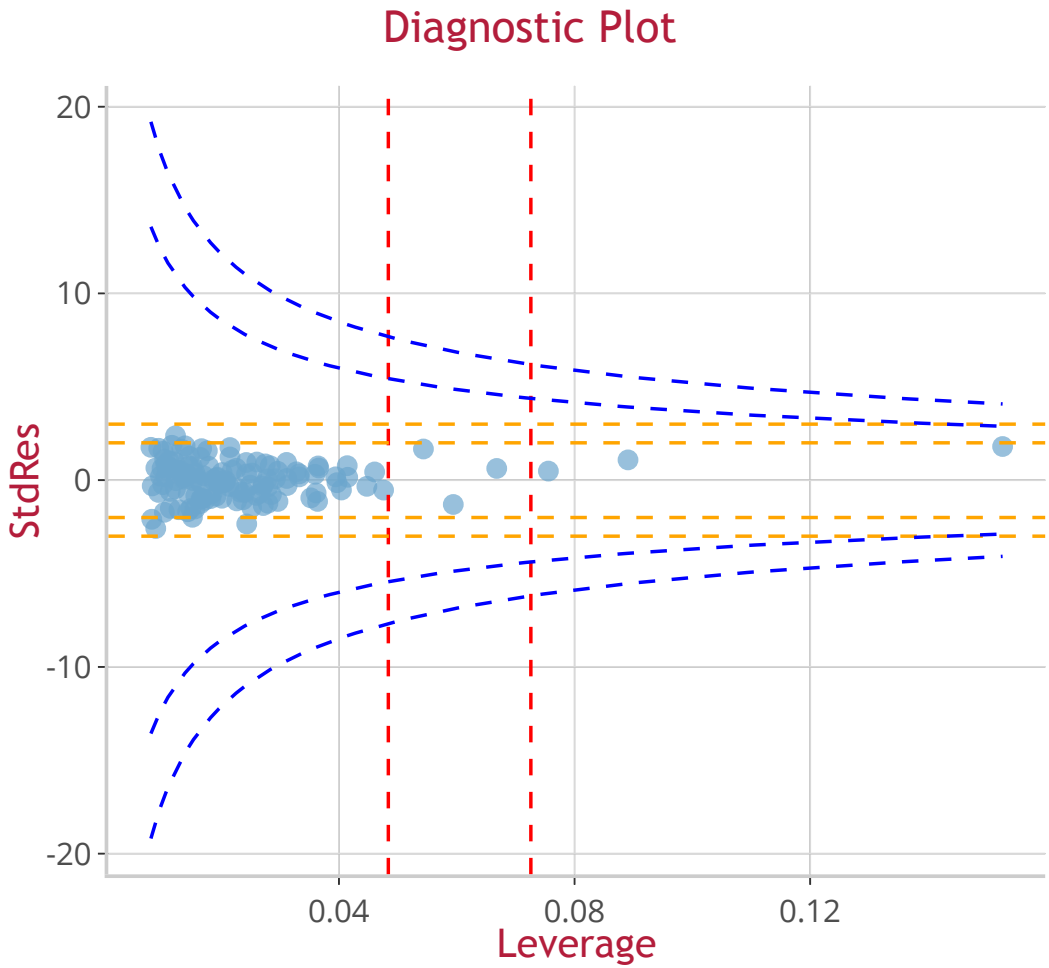
In MLR, the methods used to find unusual points are computed very similarly as those in SLR. The cutoffs are the same except for leverage, the two cutoffs are $> 2(K+1)/n$ and $3(K+1)/n$, where K is number of predictors in the model.

Statistic	Moderately unusual	Very unusual
Leverage, h_i	$> 2(K+1)/n$	$> 3(K+1)/n$
Standardized residual, stdres$_i$	beyond ± 2	beyond ± 3
Studentized residual, studres$_i$	beyond ± 2	beyond ± 3
Cook's distance, D_i	> 0.5	> 1

Unusual points

```
library(MASS)
HP.complete <- HappyPlanet[complete.cases(HappyPlanet), ] # complete dataset
n <- nrow(HP.complete) # sample size
K <- 2 # number of predictors
Leverage <- hatvalues(m3) # leverage values
StdRes <- stdres(m3) # standardized residuals
CooksD <- cooks.distance(m3) # Cook's D
unusual <- data.frame(HP.complete, Leverage, StdRes, CooksD)
cd <- function(h, type){sqrt((1-h)/h)*type} # function for cook's D cutoffs
diagplot <- ggplot(unusual, aes(x=Leverage, y=StdRes, label=Country))+
  geom_point(color="skyblue3", size=2.5, alpha=0.7)+
  geom_vline(xintercept = c(6/n, 9/n), color="red", linetype=2)+
  geom_hline(yintercept = c(-3,-2,2,3), color="orange", linetype=2)+
  stat_function(fun=cd, args=list(type=sqrt(K+1)), color="blue", linetype=2)+
  stat_function(fun=cd, args=list(type=-sqrt(K+1)), color="blue", linetype=2)+
  stat_function(fun=cd, args=list(type=sqrt(0.5*(K+1))), color="blue", linetype=2)+
  stat_function(fun=cd, args=list(type=-sqrt(0.5*(K+1))), color="blue", linetype=2)+
  ggtitle("Diagnostic Plot")
diagplot
```

Unusual points



- There are several points with unusually large leverage values and somewhat large residuals. Other than that, no point is identified by Cook's distance as unusual.

Predictions

Estimated regression line

$$\widehat{Happiness} = 25.6 - 5.7 \times log(GDPpc) + 1.0 \times LifeExp$$

1. What is the mean *Happiness* score for countries with *GDPpc* \$380 and *LifeExp* 55?
2. What is the *Happiness* score for a country with *GDPpc* \$380 and *LifeExp* 55?
3. What is the mean *Happiness* score for countries with *GDPpc* \$48,000 and *LifeExp* 77?
4. What is the *Happiness* score for a country with *GDPpc* \$48,000 and *LifeExp* 77?

- Mean *Happiness* score with 95% confidence interval
- Individual *Happiness* score with 95% prediction interval

Predictions

```
predict(m3, list(GDPpc=380, LifeExp=55), interval="confidence")
```

```
##        fit      lwr      upr
## 1 44.93723 42.12574 47.74873
```

```
predict(m3, list(GDPpc=380, LifeExp=55), interval="prediction")
```

```
##        fit      lwr      upr
## 1 44.93723 28.12785 61.74662
```

- For countries with GDPpc $380 and life expectancy 55, the average happiness score is predicted as 44.9. We are 95% confident that the interval [42.1, 47.7] will contain the true population mean response.
- For a country with GDPpc $380 and life expectancy 55, the happiness score is predicted as 44.9. We are 95% confident that the interval [28.1, 61.7] will contain the true population individual response.
- In 2006, Madagascar had *GDPpc* $379.1, *LifeExp* 55.4 and *Happiness* score 46.0.

Predictions

```
predict(m3, list(GDPpc=48000, LifeExp=77), interval="confidence")
```

```
##        fit      lwr      upr
## 1 38.65434 35.86335 41.44533
```

```
predict(m3, list(GDPpc=48000, LifeExp=77), interval="prediction")
```

```
##        fit      lwr      upr
## 1 38.65434 21.84837 55.46031
```

- For countries with GDPpc $48,000 and life expectancy 77, the average happiness score is predicted as 38.7. We are 95% confident that the interval [35.9, 41.4] will contain the true population mean response.
- For a country with GDPpc $48,000 and life expectancy 77, the happiness score is predicted as 44.9. We are 95% confident that the interval [21.8, 55.5] will contain the true population individual response.
- In 2006, United States had *GDPpc* $48,061.5, *LifeExp* 77.4 and *Happiness* score 28.8.

Summary

- **CHOOSE**
  - Exploratory data analysis
  - Model definition $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \epsilon$, where $\epsilon \overset{iid}{\sim} N(0, \sigma)$
- **FIT**: Maximum likelihood estimation (MLE)
- **ASSESS model**
  - Inference for the intercept and slopes
- **ASSESS error**: Check conditions; unusual points
- **USE**: Predictions