ME EN 2550

Ryan Dalby u0848407

Homework 7- Part A

7.1: A residual plot versus x of (a) would look like randomly distributed values around the line of 0 with steady variance, this is due to the uniform nature of the data in relation to the regression line. Looking at (b) we would have values still centered around the line of 0 but with much larger variability for smaller x values and as x gets larger the variability gets much smaller but approaches being above the line 0.

7.12: a) The height and volume of the timber have a weak positive relationship. Possibility linear but hard to tell. The relationship is weak since the data is very dispersed about the plot.

b) The diameter and volume of the timber have a strong positive relationship that appears linear. This is evident because the variability of residuals is relatively consistent and small if a linear line is drawn through the data.

c) The diameter measurement of the tree would seem to be much more predictive of the volume of the tree in a linear regression model. This is clear because of the strong linear relationship between diameter and volume and compared to height and volume is much less dispersed.

7.24: a) The relationship between calories and carbs is a positive, weak, and possibly linear one. The residuals and data are very dispersed, shown by the fanned out residual plot.

b) Explanatory: Calories Response: Carbs

c) In order to predict the number of carbs given a known calorie value.

d) No, this data does not meet the conditions required for fitting a least squares line. This is because we do not have constant variability. It does approximately meet linearity and normality conditions.

7.25: a) $\hat{y} = 50.6 + 0.726x$

b) For each mile, this model predicts a .726 minute increase in travel time. When distance traveled is 0 miles the predicted travel time is 51 minutes. (Doesn't make much sense in context)

c) $R^2 = 0.4045$ This means that 40.45% of the variability in travel time is explained by the model.

d) $\hat{y}(x=103\text{miles}) = 125.38$ minutes

e) residual = 168 - 125.38 = 42.62 minutes This means we underestimated the actual value since we got a positive value.

f) No, it would not be appropriate to use this model since it's range of validity is up to approximately 400 miles (The range of data we have). If we were to do so it would be extrapolation.

7.28: a) R = -0.8485 (Negative since the relationship on the scatterplot appears negative)

b) slope: -0.5371 intercept: 0.5534

c) When 0% receive reduced-fee lunch we expect 55.34% of bike riders to wear helmets.

d) For every percent increase in the average percentage of students receiving reduced-fee lunch, we expect the percentage wearing helmets to decrease by -0.5371 percent.

e) Expected: $\hat{y} = 0.5534 - 0.5371(.40) = 0.3386 = 33.86\%$ Actual: 40%

Residual: 40-33.86 = 0.0614 This means we underestimated the percentage wearing helmets for this given neighborhood since we have positive residual.

7.30: a) $\hat{y} = -0.357 + 4.034x$ where x is body weight and y is heart weight

b) When body weight is 0kg we predict a heart weight of -0.357g.

c) For every 1kg increase in body weight, we predict a 4.034g increase in heart weight.

d) 64.66% of the variability of heart weight is explained by the model.

e) R = +0.8041

7.31: (a) has an outlier on the bottom right of the plot, this is a high leverage point as it is far away and is an influential point since the line is greatly influenced by this outlier.

(b) has an outlier on the bottom right as well, although it is high leverage it does not end up being an influential point as it does not impact the line that much.

(c) has an outlier at the top middle. This point is not a high leverage point and is not an influential point either as it lies within the primary domain of data.

7.36: a) There is a moderately strong, positive, linear relationship between the number of cans of beer and BAC.

b) $\hat{y}$ = -0.0127 + 0.0180x where x is cans of beer and y is BAC. When 0 cans of beer have been consumed we predict a -.0127 BAC. For each unit increase in cans of beer, we expect a 0.0180 increase in BAC.

c) Ho: beta1 = 0 Ha: beta1 ≠ 0 Where beta is the actual predicted change in BAC for a unit increase of cans of beer.

P-value = approximately 0 Thus we reject the null hypothesis P-value < alpha = 0.05.

d) R^2 = 0.7921 This means that 79.21% of the variability in BAC is explained by the model.

e) It is likely the relationship would be similarly as strong but it is hard to know for sure because we would likely be looking at a different population.

7.37: a) Ho: beta1 = 0 Ha: beta1 > 0 Where beta is the actual predicted increase in wife's height given a unit increase of husbands height.

P-value = 0.0000 Since P-value < alpha = 0.05 we reject the null hypothesis(even though the test is two-sided in the table it is very small) and accept the null hypothesis meaning there is strong evidence that there is some relationship between a husband's height and a wife's height.

b) $\hat{y}$ = 43.5755 + 0.2863x where x is husband's height and y is wife's height.

c) When a husbands height is 0in we expect their wife's height to be 43.5755in. Given a unit increase in husband's height, we expect an increase of 0.2863in to their wife's height.

d) R = +0.30

e) $\hat{y}(x = 69in) = 43.5755 + 0.2863(69in) = 63.33in$. R^2 is 0.09 which is quite low so this prediction is not very reliable.

f) It is not wise to use the same linear model because 79 inches is not in the domain of data and thus using the model would be an interpolation.

7.44: a) Ho: $\beta_1 = 0$ Ha: $\beta_1 > 0$ Where beta is the actual predicted increase in heart weight given a unit increase in body weight.

b) P-value = 0.000 Since P-value < alpha = 0.05 so we reject the null hypothesis and conclude there is a positive association between body weight and heart size.

c) 95%CI: $3.544 < \beta_1 < 4.524$ This means that there is a 95% chance that the confidence interval given contains the true value of beta1 which predicts the change in heart weight of cats given a unit increase in body weight.

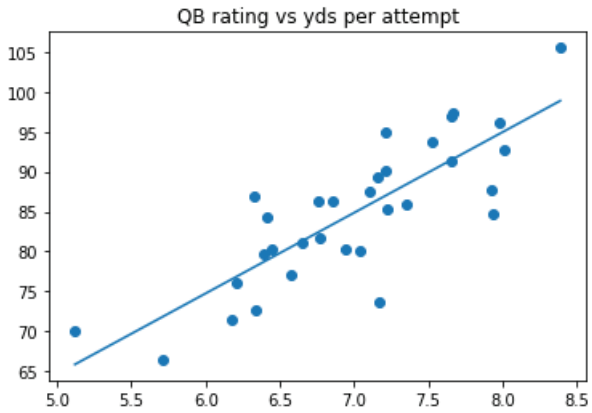
d) Yes, the results from the hypothesis test and the confidence interval agree since the value of 0 which is our null hypothesis for beta1 does not fall in the calculated confidence interval.

```
Python 3.6.5 |Anaconda, Inc.| (default, Mar 29 2018, 13:32:41) [MSC v.1900 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 6.4.0 -- An enhanced Interactive Python.

In [1]: runfile('C:/Users/hoops/OneDrive/Documents/School/ME EN 2550 Statistics and
Probability/HW7/HW7.py', wdir='C:/Users/hoops/OneDrive/Documents/School/ME EN 2550
Statistics and Probability/HW7')
```

B1:



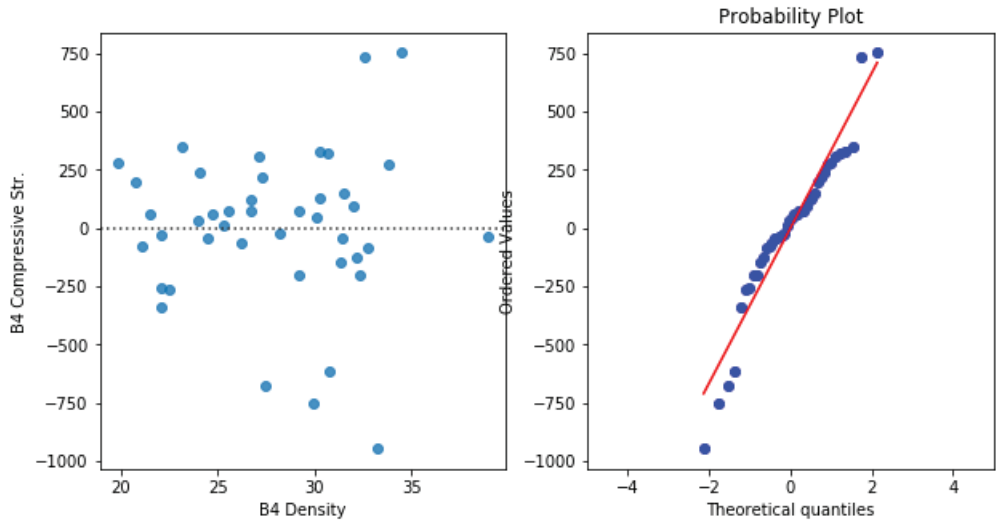
a) Slope = 10.09174 Intercept = 14.19549
b) There is a predicted 10.1 increase in QB rating given a unit increase in yards per attempt. When yards per attempt is 0 there is a predicted QB rating of 14.2
c) To increase the mean rating by 10 points yards per pass attempt should be increased by 0.99091
d) With x = 7.21 yds/attempt a QB rating of 86.95697 is predicted

B2:
a) Slope = 3.32437 Intercept = 13.32018
b) The predicted selling price given that the taxes paid are x = 7.50 is predicted to be 38.25296
c) When taxes paid is x = 5.898 the actual value is 30.90000 and the predicted value is 32.92732 and the residual is -2.02732

d) Yes the plot indicates that taxes paid is a relatively effective regressor variable in predicting selling price since we see an approximately linear relationship between sale price and annual taxes, we could determine the effectiveness more by getting more data.

B3:
a) The two-sided p-value for for a hypothesis test with a null hypothesis of slope esitimate for B1 data of 0 is 9.58903e-09. Since this is less than alpha = 0.01 we reject the null hypothesis thus there is some relationship between the quarterback rating and yards per attempt.
b) The estimated standard error of the slope is 1.28781 and the estimated standard error of the intercept is 9.05898

B4:
a) Slope = 184.55281 Intercept = -2149.64861
b) The two-sided p-value for for a hypothesis test with a null hypothesis of slope esitimate for B4 data of 0 is 1.17207e-18. Since this is less than alpha = 0.05 we reject the null hypothesis thus there is some relationship between density and compressive strength
c) R^2 = 0.85972 and thus 85.972% of the variance of compressive strength is explained by the model
d) From the plots we can see that our data is approximately normal(residual plot data is approximately flat) has nearly normal residuals(from probability plot of residuals) with constant variability based on the plots and thus conditions for doing least squares regression are met.



In [2]:

```
# -*- coding: utf-8 -*-
"""
Created on Wed Apr  3 21:18:23 2019
HW 7
@author: Ryan Dalby
"""

import pandas as pd
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt
from pyfinance.ols import OLS
import seaborn as sb

print("B1:")

data = pd.read_csv("HW7Data.csv")
yds = data['B1 Yds per Attempt'].dropna() #x like
rating = data['B1 Rating'].dropna() #y like
b1slope, b1intercept, _, b3Pval, b3SlopeSE = stats.linregress(yds, rating)
b1LSRL = lambda x: b1slope * x + b1intercept
plt.scatter(yds, rating)
ydsVals = np.linspace(np.min(yds), np.max(yds), 100)
plt.plot(ydsVals, b1LSRL(ydsVals))
plt.title("QB rating vs yds per attempt")
plt.show()
print("a) Slope = {:.5f} Intercept = {:.5f}".format(b1slope, b1intercept))

print("b) There is a predicted 10.1 increase in QB rating given a unit increase in yards per attemp


print("c) To increase the mean rating by 10 points yards per pass attempt should be increased by {:

print("d) With x = 7.21 yds/attempt a QB rating of {:.5f} is predicted".format(b1LSRL(7.21)))


print("\n\n")
print("B2:")

annualTaxes = data['B2 Taxes'].dropna() #x like
salePrice = data['B2 Sale Price'].dropna() #y like
b2slope, b2intercept, _, _, _ = stats.linregress(annualTaxes, salePrice)
b2LSRL = lambda x: b2slope * x + b2intercept

print("a) Slope = {:.5f} Intercept = {:.5f}".format(b2slope, b2intercept))

print("b) The predicted selling price given that the taxes paid are x = 7.50 is predicted to be {:.

b2PredictedValAtPoint = b2LSRL(5.898)
b2ActualValAtPoint = salePrice[np.where(annualTaxes == 5.898)[0][0]]
b2Residual = b2ActualValAtPoint- b2PredictedValAtPoint
print("c) When taxes paid is x = 5.898 the actual value is {:.5f} and the predicted value is {:.5f}

plt.scatter(annualTaxes, salePrice)
annualTaxesVals = np.linspace(np.min(annualTaxes), np.max(annualTaxes), 100)
plt.plot(annualTaxesVals, b2LSRL(annualTaxesVals))
```

```
plt.title("sale price vs annual taxes")
plt.show()
print("d) Yes the plot indicates that taxes paid is a relatively effective regressor variable in pr


print("\n\n")
print("B3:")

print("a) The two-sided p-value for for a hypothesis test with a null hypothesis of slope esitimate

b1model = OLS(rating, yds)
b3InterceptSE = b1model.se_alpha
print("b) The estimated standard error of the slope is {:.5f} and the estimated standard error of t


print("\n\n")
print("B4:")

density = data['B4 Density'].dropna() #x like
compStr = data['B4 Compressive Str.'].dropna() #y like
b4slope, b4intercept, b4CorrCoeff, b4Pval, _ = stats.linregress(density, compStr)
b4LSRL = lambda x: b4slope * x + b4intercept
print("a) Slope = {:.5f} Intercept = {:.5f}".format(b4slope, b4intercept))

print("b) The two-sided p-value for for a hypothesis test with a null hypothesis of slope esitimate

print("c) R^2 = {:.5f} and thus {:.3f}% of the variance of compressive strength is explained by the

b4fig, (b4ax1, b4ax2) = plt.subplots(ncols=2, figsize = (10,5))
sb.residplot(density, compStr, ax = b4ax1)
predictedCompStr = b4LSRL(density)
b4ResidualData = compStr - predictedCompStr
stats.probplot(b4ResidualData, plot = b4ax2)
b4ax2.set_xbound(-5, 5)
print("d) From the plots we can see that our data is approximately normal(residual plot data is app
```