Answer 11.89

We perform regression in stata:

```
. regress incidence time_period
```

Source	SS	df	MS	Number of obs	=	5
				F(1, 3)	=	19.09
Model	11262.736	1	11262.736	Prob > F	=	0.0222
Residual	1770.292	3	590.097333	R-squared	=	0.8642
				Adj R-squared	=	0.8189
Total	13033.028	4	3258.257	Root MSE	=	24.292

incidence	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
time_period	33.56	7.681779	4.37	0.022	9.11315	58.00685
_cons	184.9	25.47758	7.26	0.005	103.819	265.981

The regression line is $y = 184.9 + 33.56x$.

Answer 11.90

The hypothesis is: $H_0: \beta = 0$, $H_1: \beta \neq 0$. We use the F test for simple linear regression.

Since $F = 19.09$ and $\Pr(F > 19.09) = 0.0222 < 0.05$. We can conclude that there is a significant association between time period and diabetes incidence with incidence increasing over the past 25 years.

Answer 11.95

We import the data into Excel spreadsheets and compute the rank of weight change and A1C change using RANK.AVG function.

	id	WtD	A1cD	WtRank	A1cRank
1	1	5	-1.5	9	4
2	2	3.8	-2.1	6	1
3	3	5.7	-.8	11	6
4	4	4.5	.7	8	13
5	5	3.3	-1.9	5	2.5
6	6	6.4	-.8	13	6
7	7	.9	.4	3	10
8	8	.6	.6	2	12
9	9	-.2	1.8	1	15
10	10	3.2	.8	4	14
11	11	5.6	0	10	8
12	12	4.3	.5	7	11
13	13	6	.3	12	9
14	14	7.2	-.8	14	6
15	15	7.9	-1.9	15	2.5

We use Pearson correlation in stata:

```
. pwcorr WtRank A1cRank, sig

             |   WtRank  A1cRank
-------------+------------------
      WtRank |   1.0000
             |
             |
     A1cRank |  -0.5238   1.0000
             |   0.0451
             |
```

The correlation coefficient is -0.5238. The p-value is 0.0451.

```
. ci2 WtRank A1cRank, corr

Confidence interval for Pearson's product-moment correlation
of WtRank and A1cRank, based on Fisher's transformation.
Correlation = -0.524 on 15 observations (95% CI: -0.817 to -0.016)
```

The 95% confidence interval for the true rank correlation is (-0.817, -0.016). We conclude that there is a negative relationship between weight change and A1C change.

Answer Additional Problem

Data analysis problems:

1) Use the FEV data from Rosner to answer the research question: Is there a relationship between smoking status and FEV after adjusting for Height, Sex, and age.

a. Categorize age by every 5 years [1-5, 6-10,....].

We generate a variable cate_age and use it to categorize age using the following commands.

```
. generate cate_age=.
(654 missing values generated)

. replace cate_age=1 if Age >1 & Age <=5
(39 real changes made)

. replace cate_age=2 if Age>5 & Age <=10
(351 real changes made)

. replace cate_age=3 if Age>10 & Age <=15
(234 real changes made)

. replace cate_age=4 if Age>15 & Age <=20
(30 real changes made)
```

b. Fit the preliminary model and assess assumptions. Discuss any violated assumptions if present, or determine they all hold.

We fit the model: $Fev = \alpha + \beta_1 \times Smoking + \beta_2 \times height + \beta_3 \times Sex + \beta_4 \times cate\_age + e$

```
. regress fev Smoke Hgt Sex cate_age
```

Source	SS	df	MS		
				Number of obs	= 654
				F(4, 649)	= 549.75
Model	379.050088	4	94.762522	Prob > F	= 0.0000
Residual	111.869748	649	.172372493	R-squared	= 0.7721
				Adj R-squared	= 0.7707
Total	490.919836	653	.75179148	Root MSE	= .41518

fev	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
Smoke	-.0438879	.0583888	-0.75	0.453	-.1585417	.0707659
Hgt	.1118022	.004196	26.64	0.000	.1035628	.1200417
Sex	.1511428	.0333881	4.53	0.000	.0855811	.2167044
cate_age	.2215195	.0360479	6.15	0.000	.150735	.292304
_cons	-4.801909	.2030969	-23.64	0.000	-5.200716	-4.403103

We then conduct residual analysis for the model.

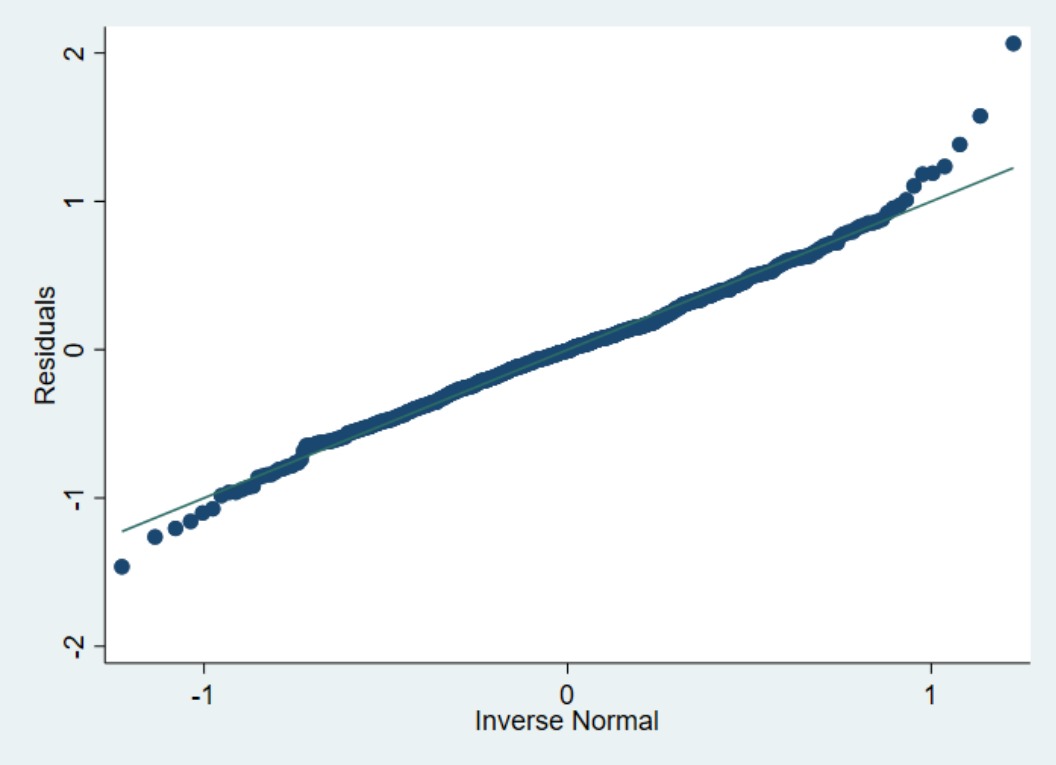

Figure 1. QQ plots of raw residuals.

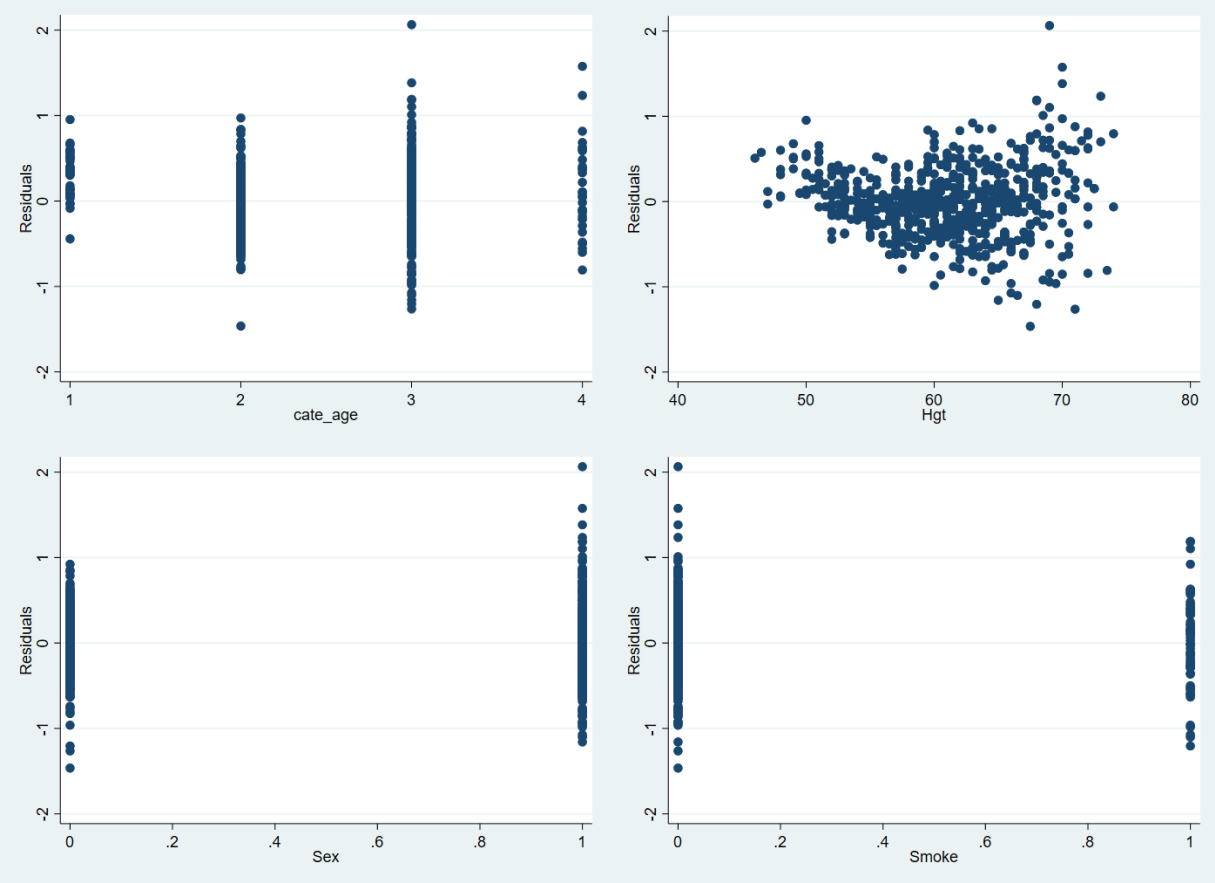

Figure 2. Scatter plots of raw residuals.

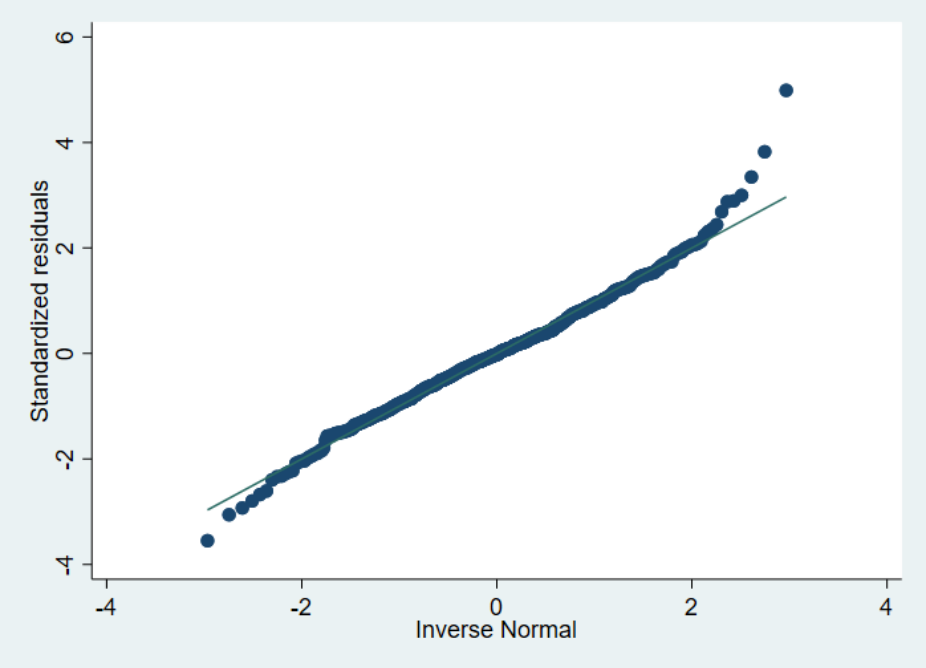

Figure 3. QQ plot of standardized residuals.

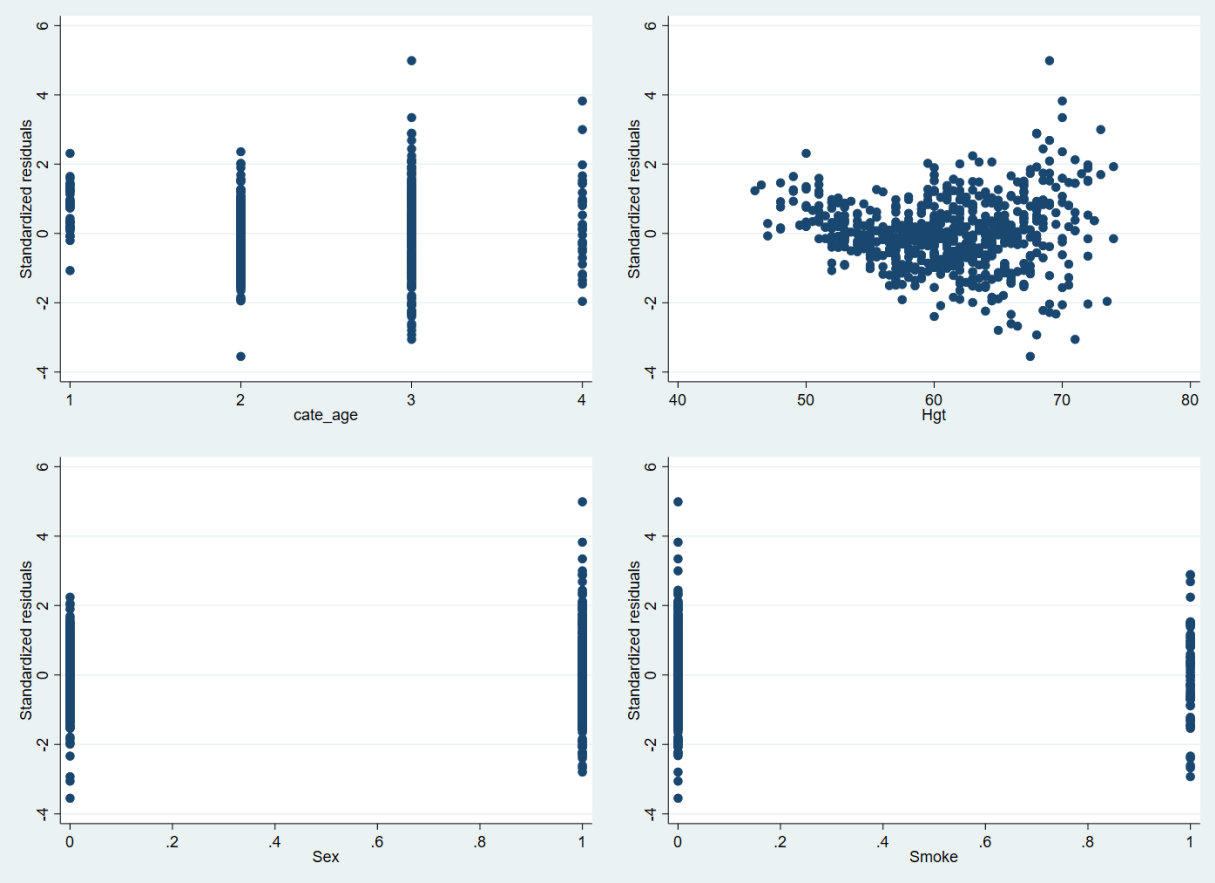

Figure 4. Scatter plot of standardized residuals.

There are slight deviations from the diagonal in the QQ plots. Therefore, the preliminary model does not fit very well.

c. Using the gladder command, which transformation of FEV would you choose? Explain why in 2 or 3 sentences.

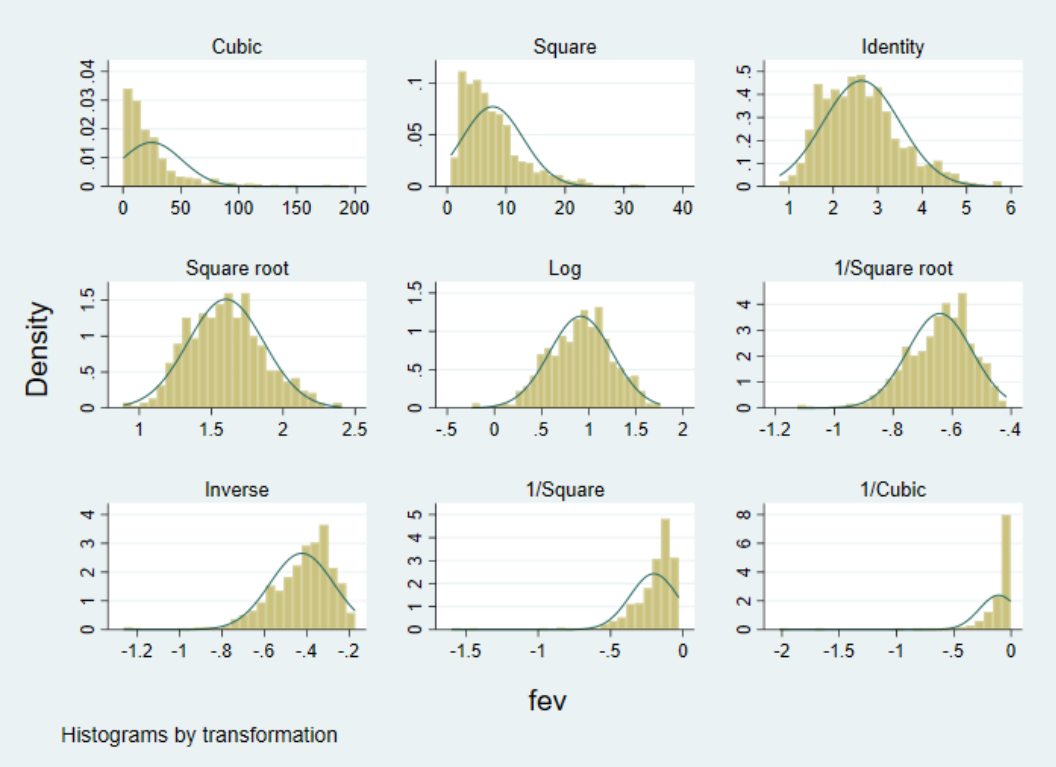

Figure 5. Histograms of transformed FEV.

The histogram shows that log(Fev) fits the normal distribution well. Therefore, we try transform FEV with log(Fev).

d. For uniformity of answers on this homework, fit a second model using log(Fev), and assess the assumptions.

We fit the model: $\log(Fev) = \alpha + \beta_1 \times Smoking + \beta_2 \times height + \beta_3 \times Sex + \beta_4 \times cate\_age + e$

```
. regress logfev Smoke Hgt Sex cate_age
```

Source	SS	df	MS		
Model	58.54218	4	14.635545	Number of obs	= 654
Residual	13.9837347	649	.021546587	F(4, 649)	= 679.25
				Prob > F	= 0.0000
				R-squared	= 0.8072
				Adj R-squared	= 0.8060
Total	72.5259147	653	.111065719	Root MSE	= .14679

logfev	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
Smoke	-.0297791	.0206436	-1.44	0.150	-.0703153	.0107572
Hgt	.0456956	.0014835	30.80	0.000	.0427825	.0486087
Sex	.0269478	.0118045	2.28	0.023	.0037683	.0501274
cate_age	.0767864	.0127449	6.02	0.000	.0517602	.1018125
_cons	-2.072952	.0718057	-28.87	0.000	-2.213951	-1.931952

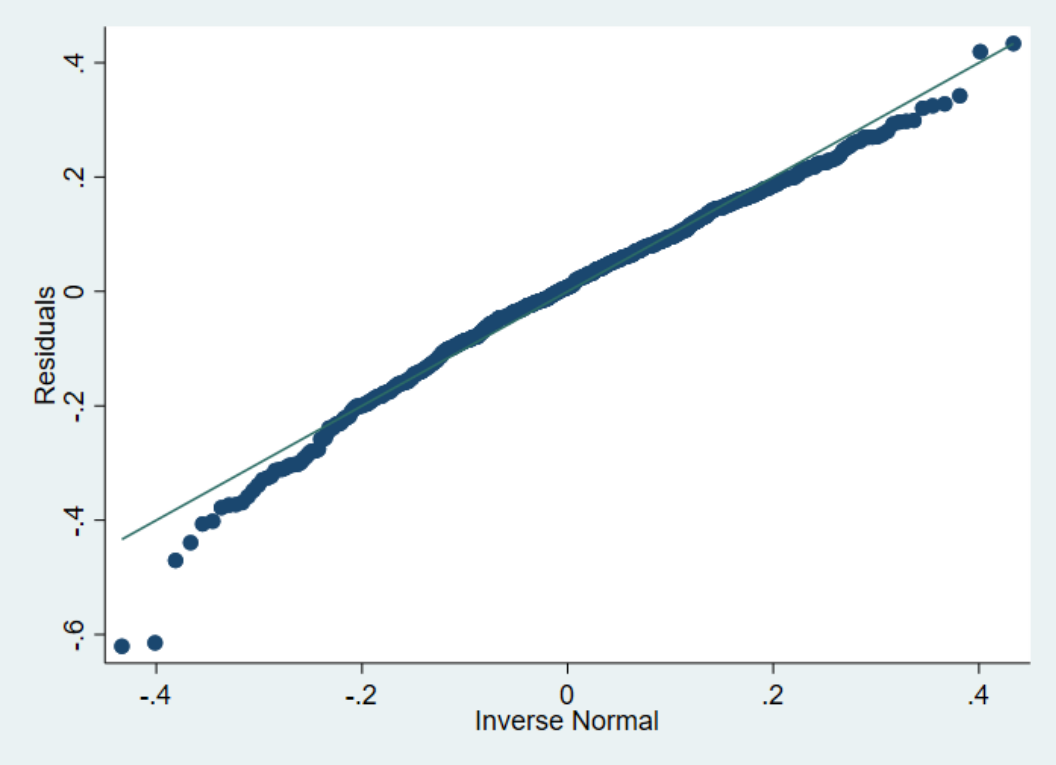

Figure 5. QQ plot of raw residuals.

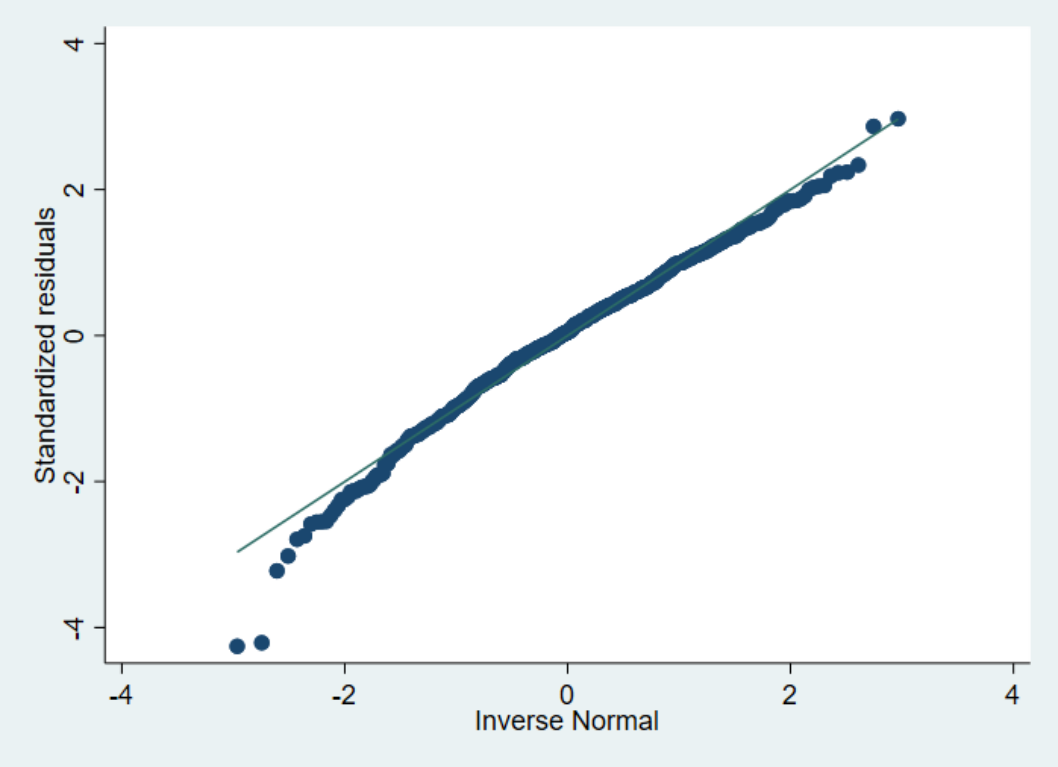

Figure 6. QQ plot of standardized residuals.

There are slight deviations from the diagonal in the QQ plots. But overall the regression model fits.

e. Using the model from part d, answer the following questions

i. Is the model significant? Report the test statistic and p-value used for answering the question.

The model is significate, since F=679.25 and the p-value is 0.00<0.05.

ii. How much variation of log(Fev) is explained by the model?

The R-squared value is 0.8072. Therefore 80.72% of variation can be explained by the model.

iii. Is the coefficient for Smoking significant? (Yes or no) What is the 95% confidence interval for the coefficient, and how is it interpreted?

No, since the p-value is 0.15>0.05. The confidence level of the coefficient is (-0.07, 0.01). This implies that we are 95% confident that the value of the coefficient is between -0.07 and 0.01.

iv. Is the coefficient for Height significant? How would we interpret it?

Yes, since the p-value is 0.00<0.05. This means an increase of 1 inch in height will result in an increase of 0.0457 in log(Fev) while other variables remain constant.

v. Are the categories of Age significant? Report the hypothesis, test statistic and p-value. How would we interpret the coefficient for the oldest category?

Yes, since the p-value is 0.00<0.05. The hypothesis is: $H_0: \beta_4 = 0$, $H_1: \beta_4 \neq 0$. The test statistic is 6.02. This means if the person is in older age category log(Fev) will increase 0.0768 while other variables remain constant.

f. Using the model, predict the log(FEV) for an average 17 year old male who is 60 inches tall and not a current smoker. What would the predicted FEV be?

$$\log(Fev) = -0.0298 \times 0 + 0.0457 \times 60 + 0.0269 \times 1 + 0.0768 \times 4 - 2.07 = 1.0061$$