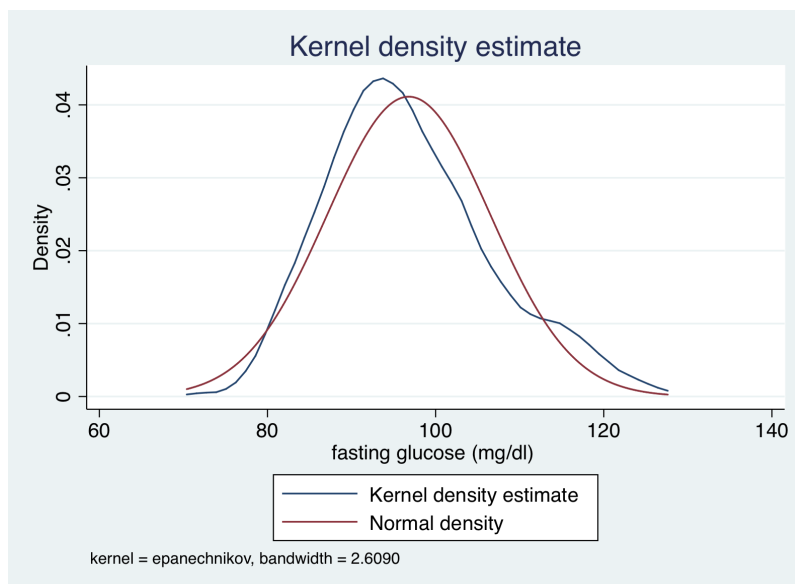
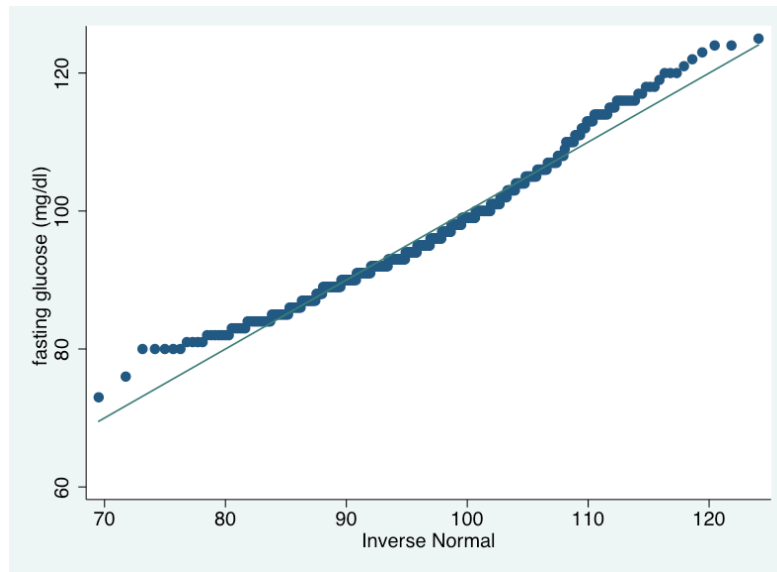
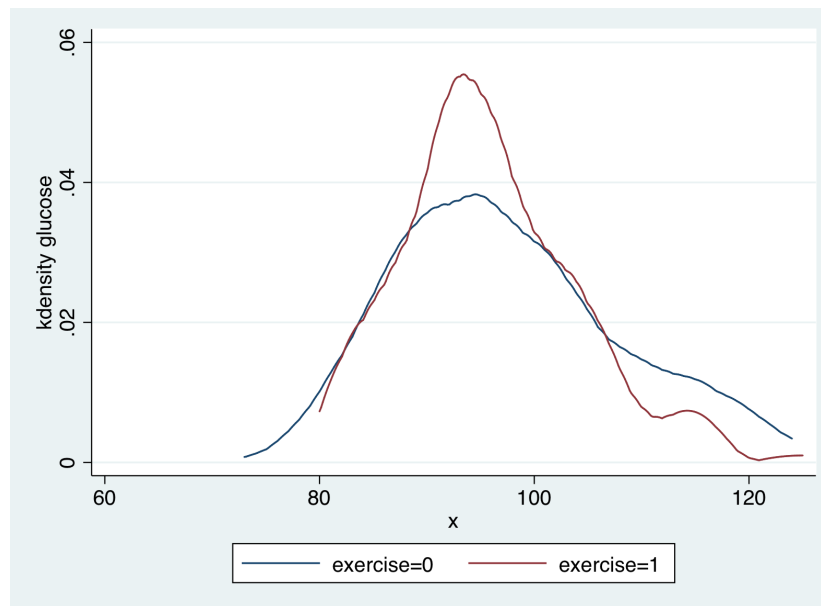


Comments on Biostatistics 208 Lab #2 1/16/20

The Q-Q plot of glucose shows a little right skewness, but in a sample of 406 observations, And with no outliers (thanks to the omission of women with diabetes), the departure from normality is probably not bad enough to merit using a log transformation. The accompanying density plot makes this a little clearer. Again, in a sample of this size, this degree of departure from symmetry doesn't look too bad.



The boxplots and kernel density estimates show slightly lower glucose among HERS women who exercise. There is also slightly less spread in that group, but not enough to worry about.



Question 1: The assumption of normality (i.e. within exercise groups) appears reasonable in this case. The side-by-side comparisons of the distributions in the two exercise groups also make the assumption of equal variance acceptable.

The between-group differences in glucose shown in the plots are reflected in the regression results:

. regress glucose exercise

Source	SS	df	MS
Model	250.226713	1	250.226713
Residual	37866.0097	404	93.7277469
Total	38116.2365	405	94.1141641

Number of obs = 406
 F(1, 404) = 2.67
 Prob > F = 0.1031
 R-squared = 0.0066
 Adj R-squared = 0.0041
 Root MSE = 9.6813

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exercise	-1.617965	.9902309	-1.63	0.103	-3.564614	.3286833
_cons	97.41667	.6098652	159.73	0.000	96.21776	98.61557

```
. lincom _cons + exercise
```

```
( 1) exercise + _cons = 0
```

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
(1)	95.7987	.7801421	122.80	0.000	94.26506	97.33235

Question 2:

- Mean glucose among women who exercise at least 3 times a week is 95.8 *mg/dL*, either from the `lincom` result or by adding 97.42 and -1.62 from the main output. The `lincom` result gives a potentially useful CI for the mean in this group, but the null hypothesis of the accompanying test – that mean glucose is zero in this group – is not of interest.
- Mean glucose in the women who exercise less than 3 times a week is estimated by the intercept, or 97.4 *mg/dL*.
- The between-group difference in mean levels is estimated by the coefficient for `exercise`, or -1.6 *mg/dL*.
- The 95% CI for the difference is -3.6 to .33 *mg/dL*.
- For the test of whether the between-group difference is zero, $t = -1.63$, $p = 0.103$. So there is a bit of evidence for slightly lower average glucose in the women who exercise at least 3 times a week.

These statistics could all be found in the *t*-test output as well.

Next we have the regression of `glucose` on `bmi`:

```
. regress glucose bmi
```

Source	SS	df	MS			
Model	1021.96128	1	1021.96128	Number of obs =	406	
Residual	37094.2752	404	91.8175128	F(1, 404) =	11.13	
Total	38116.2365	405	94.1141641	Prob > F =	0.0009	
				R-squared =	0.0268	
				Adj R-squared =	0.0244	
				Root MSE =	9.5821	

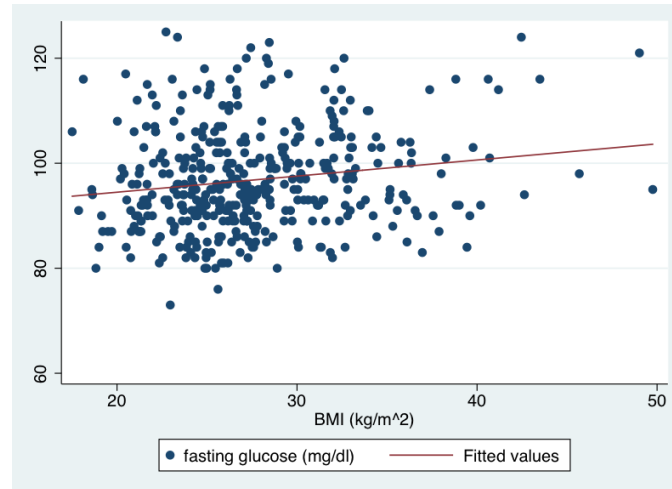
glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
bmi	.3069144	.0919947	3.34	0.001	.1260663	.4877624
_cons	88.34208	2.580269	34.24	0.000	83.26965	93.41451

Question 3:

We might summarize the association by saying that “In unadjusted analysis of data for 406 diabetes-free post-menopausal women with CHD, mean glucose levels increased 0.3 *md/dL* with

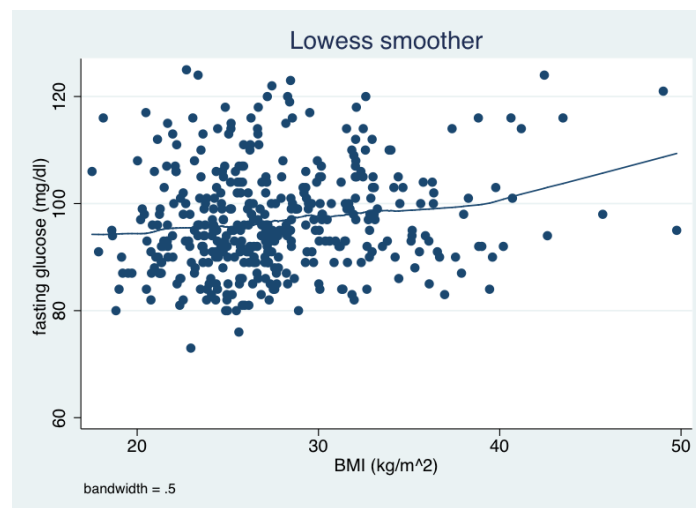
every unit increase in BMI (95% CI 0.1-0.5, $p=0.001$). However, BMI explained less than 3% of the variability in glucose levels.”

Our hands-on computation of the residual sum of squares (37094.273) agrees with the entry in Residual row and SS column of the ANOVA table in the upper left of the regression output.



The plot certainly suggests that mean fasting glucose levels get slightly but progressively higher among women with higher and higher BMI. The assumption that the increase is approximately linear in BMI appears reasonable, in view of the following Lowess plots.

Question 4: There is a slight upturn in the Lowess estimate at the right, but it appears to be driven by just a few relatively observations. Given the noisiness of these smoothers in the tails, it seems doubtful that the apparent curvature is real and the assumption of linearity appears justified.



The following models use two transformations of BMI. The first rescales the predictor, to make the units of the slope more interpretable.

```
. gen bmi5 = bmi/5  
. regress glucose bmi5
```

Source	SS	df	MS	Number of obs = 406		
Model	1021.96124	1	1021.96124	F(1, 404) = 11.13		
Residual	37094.2752	404	91.8175129	Prob > F = 0.0009		
				R-squared = 0.0268		
				Adj R-squared = 0.0244		
				Root MSE = 9.5821		
Total	38116.2365	405	94.1141641			

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
bmi5	1.534572	.4599735	3.34	0.001	.6303315	2.438812
_cons	88.34208	2.580269	34.24	0.000	83.26965	93.41451

When we substitute predictor `bmi5` for `bmi`, the coefficient estimate is inflated by a factor of exactly 5, as are the SE and 95% confidence limits; the coefficient is now interpretable as the change in mean glucose for every 5-unit increase in BMI. However, the *t*-statistic and *p*-value are unchanged (why?), as are all results for the intercept. The second transformation centers BMI, to make the intercept interpretable.

```
. egen meanbmi = mean(bmi)
. gen cbmi = bmi - meanbmi
. regress glucose cbmi
```

Source	SS	df	MS	Number of obs = 406		
Model	1021.96128	1	1021.96128	F(1, 404) = 11.13		
Residual	37094.2752	404	91.8175128	Prob > F = 0.0009		
				R-squared = 0.0268		
				Adj R-squared = 0.0244		
				Root MSE = 9.5821		
Total	38116.2365	405	94.1141641			

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
cbmi	.3069144	.0919947	3.34	0.001	.1260663	.4877624
_cons	96.80296	.4755539	203.56	0.000	95.86809	97.73782

Question 5: Centering BMI has no effect on any of the statistics for that predictor, but changes everything for the intercept. In the two first models, the intercept estimates mean glucose levels among women with BMI of zero, which is not interpretable, while in the model using `cbmi`, the intercept estimates mean glucose levels among women with BMI at the sample average of approximately 27.6 kg/m^2 .

In the model using `stdbmi`, the *t*-statistic and *p*-value for BMI again remain unchanged. This is expected, because re-scaling does not affect the overall inferences about the effects of this variable. However, re-scaling does change the coefficient estimate, standard error, and confidence interval. The coefficient estimate of 1.56 is interpretable as the increase in mean glucose, in *mg/dL*, for every standard deviation ($\approx 5.2 \text{ kg/m}^2$) increase in BMI – and is close, for obvious reasons, to the result using `bmi5`. The intercept is the same as in the model using `cbmi`, and estimates mean glucose levels in women with sample average BMI.

```
. egen stdbmi = std(bmi)
```

```
. sum stdbmi
```

Variable	Obs	Mean	Std. Dev.	Min	Max
stdbmi	406	-2.74e-10	1	-1.943207	4.287785

```
. reg glucose stdbmi
```

Source	SS	df	MS	Number of obs =	406
Model	1021.96127	1	1021.96127	F(1, 404) =	11.13
Residual	37094.2752	404	91.8175128	Prob > F =	0.0009
				R-squared =	0.0268
				Adj R-squared =	0.0244
Total	38116.2365	405	94.1141641	Root MSE =	9.5821

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
stdbmi	1.588509	.4761406	3.34	0.001	.6524865 2.524532
_cons	96.80296	.4755539	203.56	0.000	95.86809 97.73782

Question 6: The gamma distribution for residuals chosen in the program code is highly skewed and is an example of the exponential distribution. You should have seen a marked improvement in the symmetry of the distribution of β_1 with increasing sample size.