Problem 1

a) Follow instructions.

b) Stata Command: infix year 8-11 tenure 77-79 hrlyyearn 83-88 sex 21 industry 33-34 using pub1116.prn, clear

c) Hourly wages are measured with two decimal places, therefore if you divide your hourly wage variable by 100 you will get the more familiar dollar values of a worker's wage. The detailed summary of hourly wages, and tenure (measured in months) is shown below:

hrlyyearn

	Percentiles	Smallest		
1%	10	2.08		
5%	11	3		
10%	11.75	3.03	Obs	51642
25%	15	3.03	Sum of Wgt.	51642
50%	22		Mean	25.44478
		Largest	Std. Dev.	13.31136
75%	32.4	128.21		
90%	43.27	132.21	Variance	177.1924
95%	50.21	138.22	Skewness	1.414454
99%	69.23	139.42	Kurtosis	5.968722

tenure

	Percentiles	Smallest		
1%	1	1		
5%	3	1		
10%	6	1	Obs	60798
25%	20	1	Sum of Wgt.	60798
50%	66		Mean	94.75136
		Largest	Std. Dev.	84.14073
75%	163	240		
90%	240	240	Variance	7079.663
95%	240	240	Skewness	.6213356
99%	240	240	Kurtosis	1.924587

A couple things to note: the tenure variable gives a value of 240 if a workers has a tenure longer than 240 months (this could be important), and also we see that there are a different number of observations for the wage variable versus the tenure variable. To get a rough idea of why this is the case, you should notice that if an observation does not have a value for tenure (=. in Stata) then there is no value for wage; however the converse is not true, some people have no value for wage but have a value for tenure. This could occur for people who are still employed (positive tenure) but are on leave for some reason (so not earning money). We would have to dig further into the data by looking at more variables (on the observations who have no wage but do have tenure) to confirm this, but I think that is one plausible explanation.

Stata Command: sum hrlyyearn tenure, detail

d) I do a boring test of differences in mean job tenure between males and females. Hopefully you did something more interesting!

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
1	31504	96.67293	.4798377	85.16815	95.73243	97.61343
2	29294	92.68482	.4847855	82.97342	91.73461	93.63502
combined	60798	94.75136	.3412413	84.14073	94.08252	95.42019
diff		3.988114	.6827481		2.649926	5.326303

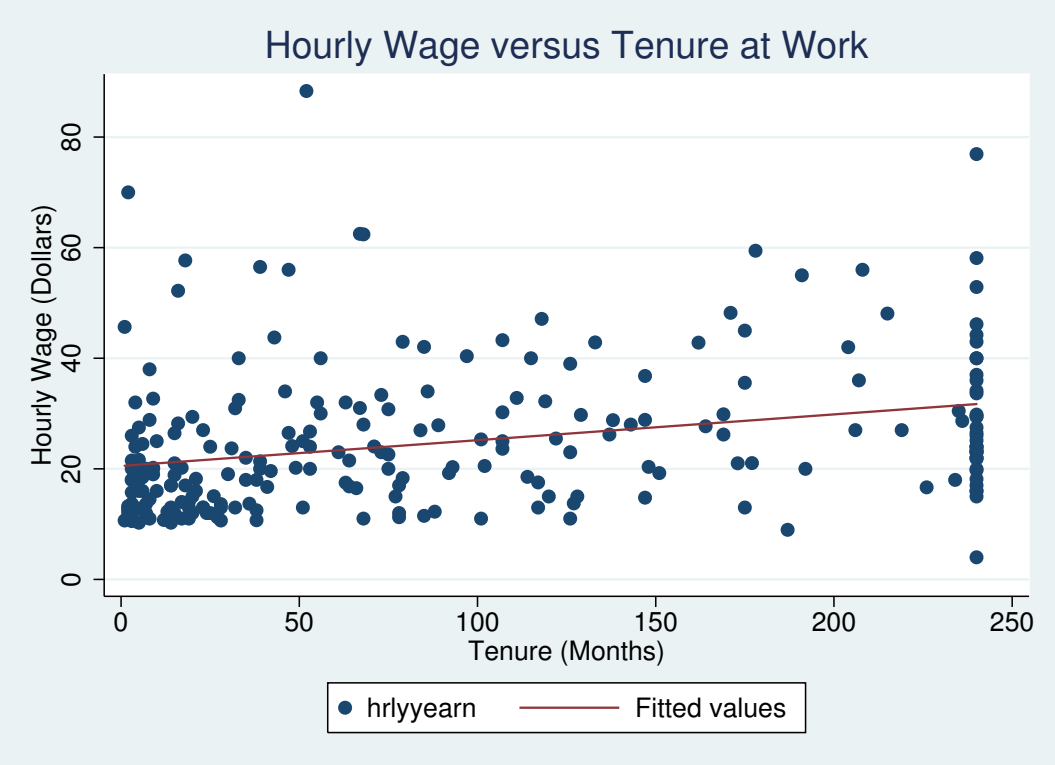
diff = mean(1) - mean(2) t = 5.8413
Ho: diff = 0 degrees of freedom = 60796

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
Pr(T < t) = 1.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 0.0000

I reject the null that men and women have equal mean tenure.

Stata Command: ttest tenure, by(sex)

e) Here is the plot of hourly wage against job tenure for only the first 500 observations.



From the line of best fit, you can see that there is a positive correlation between hourly wages and tenure.

Stata Command: twoway (scatter hrlyyearn tenure) (lfit hrlyyearn tenure) if obs≤500, ytitle(Hourly Wage (Dollars)) xtitle(Tenure (Months)) title(Hourly Wage versus Tenure at Work)

f) The population slope coefficient (call it β) is equal to $\frac{Cov(hrlyyearn,tenure)}{Var(tenure)}$. Our estimate of the slope coefficient will use the sample versions of the covariance and variance terms: therefore, $\hat{\beta} = \frac{339.685}{6710.44} = 0.0506$.

Stata Command (to get covariance/variance): corr hrlyyearn tenure, covariance.

g) Here is my regression output: we obtain the same estimate as calculated above, as expected.

Source	SS	df	MS		
Model	887968.021	1	887968.021	Number of obs	= 51642
Residual	8262422.98	51640	160.000445	F(1, 51640)	= 5549.78
				Prob > F	= 0.0000
				R-squared	= 0.0970
				Adj R-squared	= 0.0970
Total	9150391	51641	177.192367	Root MSE	= 12.649

hrlyyearn	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tenure	.0506204	.0006795	74.50	0.000	.0492886	.0519523
_cons	20.94831	.0821055	255.14	0.000	20.78739	21.10924

Stata Command: regress hrlyyearn tenure

The interpretation of this coefficient is that a one month increase in job tenure increases the conditional mean of workers' hourly wage by 5.06 cents (0.0506 dollars).[1] In other words, a one month increase in job tenure increases the expected hourly wage of this worker by 5.06 cents.

h) I would predict his hourly wage to be $20.94831 + 0.05062(60) = 23.98$. This is an estimate of the expected hourly wage of a worker *conditional on having 60 months, or 5 years, of tenure*. You derive this from the fact that our model is simply $Y_i = \alpha + X_i\beta + U_i$, and by taking the conditional expectation operator on both sides you obtain $E[Y_i|X_i = x] = \alpha + x\beta$ because $E[U_i|X_i] = 0$. So the prediction is an estimate of $E[Y_i|X = 60]$, which, given our proposed model, equals $\hat{\alpha} + 60 \cdot \hat{\beta}$, as is calculated above.[2]

i) Given the interpretation of the coefficient above, it can be seen that if we test whether or not this coefficient is zero we are testing whether or not job tenure has an effect on the mean wage for a person. As you will see later, the ratio of $\frac{\hat{\beta}}{\sqrt{Var(\hat{\beta})}}$ follows a T-distribution (or Z distribution, depending on assumptions). Thus we can do a simple T-test to determine if job tenure affects the mean wage of a worker.

[1] Technically, this interpretation is only valid assuming that the conditional expectation function is in fact linear (that is, $E[Y_i|X_i = x] = \alpha + x\beta$. If the conditional expectation is non-linear, then the interpretation changes slightly; the coefficient gives the marginal change in the best linear predictor of this conditional expectation function.

[2] This is beyond the scope of the question, but you may wonder whether predictions from a regression are of any particular significance (after all, we could use all sorts of different rules to make predictions). It turns out that the function $\alpha + x \cdot \beta$ (where α, β are regression coefficients) is the best linear predictor of y (where "best" is defined by having the smallest prediction error).