State/County Workforce Weekly Wage Evaluation

Jesse St. John and Gabriel Barela

# Research Question

How does the weekly wage affect the cost of building a single design industrial complex for a budgeted construction contract with a suitable workforce located in the continental United States?

Location Parameters:

* State
* County

**NULL Hypothesis:**

The average quarterly weekly wage is not affected by the size of the employee level availability by location.

**Alternate hypothesis:**

The average quarterly weekly wage is affected by the size of employee level availability by location.

# Method Used

Our analysis will be performed using count data methods. Count data is data in the form of whole numbers or integers; and often it is not known how often something does not occur. Thus the focus is on examining the frequencies in the data instead of looking at the proportions.

When analyzing count data, most of the conventional linear regression methods are not appropriate for use due to the following:

*Linear models may lead to negative prediction counts* Errors are not normally distributed *Variance of the response value is likely to increase with the mean* If needing to transform your data, zeros can be problematic

Specifically, we will be looking at using regression models with Poisson Errors and further investigating the data using ANOVA.

Benefits to using a Poisson errors model:

* This model uses the log link (the fitted value found using the GLM is obtained by applying the anti log of the linear predictor), which ensures all fitted values are positive
* Poisson errors account for the integer data and that the data will have variances that are equal to their means
* The deviance using a Poisson error model is calculated as follows: .

Reminder: the deviance using a linear model is calculated by:

R code:

* The glm() function is what we will use to create our Poisson regression model. We then create our model using the glm() function and specify the family argument as poisson for the odd numbered models and quasipoisson for the even numbered models.
* Summary() shows all of the information from the function for further analysis
* For the degrees of freedom we can calculate the degrees of freedom by subtracting the number of coefficients from the number of rows in our data set.

Analysis of variance (ANOVA) is used when the explanatory variables, or factors, are categorical. A one-way ANOVA can indicate if there are any statistical differences between the means of three or more independent groups. This is the version we will be using to get a closer look at our data after running the Poisson Regression. In our study, we will assume the average quarterly employee level for each state is independent. We want to see if the means in a given area(s) is significantly different or if the difference is due to chance.

# Data Set

Private, NAICS 236 Construction of buildings, All Counties 2022 Fourth Quarter, All establishment sizes Source: Quarterly Census of Employment and Wages - Bureau of Labor Statistics <http://www.bls.gov/cew/data/api/2022/4/industry/236.csv>

Additional information: <https://www.bls.gov/cew/about-data/downloadable-file-layouts/quarterly/naics-based-quarterly-layout.htm>

**QCEW Geographic Area Codes and Titles**

Quarterly Census of Employment and Wages (QCEW) data files and databases use a five character field to hold an “area code” that identifies the geographic area that the data represents. QCEW data are available at the national level, as statewide data, as county level data, and for Metropolitan Statistical Areas (MSAs) and other types of Statistical Areas. The table below shows the code and title relationships applicable to NAICS coded QCEW data. New field created to add area\_title to the data set in order to identify location name in addition to area\_fip code. For example:

* US000 = U.S. TOTAL
* USCMS = U.S. Combined Statistical Areas (combined)
* USMSA = U.S. Metropolitan Statistical Areas (combined)
* USNMS = U.S. Nonmetropolitan Area Counties (combined)
* 01000 = Alabama – Statewide
* 01001 = Autauga County, Alabama
* 01003 = Baldwin County, Alabama

**QCEW Ownership Titles (For NAICS-Based Data)**

These ownership codes are to be used with NAICS coded QCEW data. A subset of these codes are also used in the SIC based data.

The QCEW program stopped publishing data for International Government establishments as a separate ownership group after 1994. These establishments are included in private sector tabulations starting in 1995. These data were for establishments owned by international governments that reported for Unemployment Insurance purposes. Historically, there were fewer than 40 establishments, with fewer than one thousand employees coded in this group. There are published SIC coded data for this group through 1994. However, due to disclosure limitations, there are no published NAICS coded data for the period from 1990 forward.

For this analysis we will be reducing to code 5 - Private.

**QCEW Industry Codes and Titles (For NAICS Coded Data)**

The QCEW program uses the North American Industry Classification System (NAICS) as to assign establishments to industries and to report industry data at highly detailed as well as at aggregated levels. Detailed industry NAICS-coded QCEW data are available from 1990 forward, using the following time-version pairings.

For this analysis we are using NAICS 236 Construction of buildings, All Counties 2022 Fourth Quarter.

**QCEW Aggregation Level Codes**

These are the codes used on summary records produced by the QCEW program to indicate the aggregation level of the data summarized on the record. These aggregation level codes are for QCEW records coded to the North American Industry Classification System (NAICS).

For this analysis we are utilizing 55 Statewide, NAICS 3-digit – by ownership sector and 75 County, NAICS 3-digit – by ownership sector.

**Disclosure Code**

This is a 1-character disclosure code (either ’ ‘(blank) or ’N’ not disclosed). We will be reducing the data set to reference disclosed information.

**Average Quarterly Employment**

This average consists of the average estimate of workers within the three month period identified for the Q4 2022 data set.

* month1\_emplvl
* month2\_emplvl
* month3\_emplvl
* avg\_qtrly\_emplvl=(month1\_emplvl+month2\_emplvl+month3\_emplvl)/3

**Average Weekly Wage**

Average weekly wage for a given quarter.

# Identifed Variables for Analysis

Create a new data frame called dat.private.state was that reduces the column information to:

* Owner Code = 5 = Private
* Aggregation Level Code = 55 = Statewide
* Disclosure Code != “N” = Disclosed information
* Average Quarterly Employee Level > 4000
* Average Weekly Wage < 2000

Create another new data frame called dat.private.county that reduces the column information to:

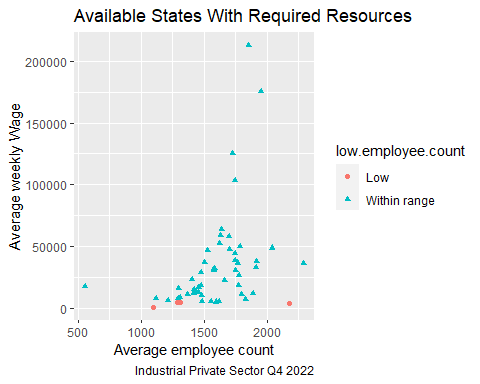
* Owner Code = 5 = Private
* Aggregation Level Code = 75 = County Level
* Disclosure Code != “N” = Disclosed information
* Average Quarterly Employee Level > 1000
* Average Weekly Wage < 2000

# File Upload

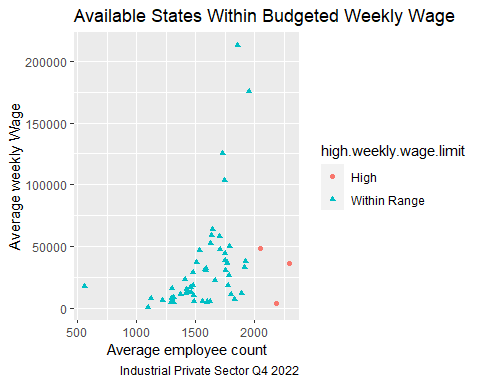
dat<-read.csv("Wage2022Q4.csv",header=TRUE)  
  
dat$total\_emplvl<- dat$month1\_emplvl + dat$month2\_emplvl + dat$month3\_emplvl  
dat$avg\_qtrly\_emplvl<-as.integer(dat$total\_emplvl / 3)

# Statewide Available Workforce by Weekly Wage Analysis

dat.private.state<-data.frame(dat[dat$own\_code == 5 & dat$agglvl\_code == 55 & dat$disclosure\_code != "N",])  
dat.private.state$low.employee.count<-ifelse(dat.private.state$avg\_qtrly\_emplvl>=5000,"Within range", "Low")  
dat.private.state$high.weekly.wage.limit<-ifelse(dat.private.state$avg\_wkly\_wage<2000,"Within Range","High")  
  
  
ggplot(data= dat.private.state, mapping = aes(x = avg\_wkly\_wage,y = avg\_qtrly\_emplvl ))+geom\_point(mapping = aes(color = low.employee.count, shape = low.employee.count ))+labs(   
 title="Available States With Required Resources",  
 caption="Industrial Private Sector Q4 2022",  
 x="Average employee count",  
 y="Average weekly Wage",)



ggplot(data= dat.private.state, mapping = aes(x = avg\_wkly\_wage,y = avg\_qtrly\_emplvl ))+geom\_point(mapping = aes(color = high.weekly.wage.limit, shape = high.weekly.wage.limit ))+labs(   
 title="Available States Within Budgeted Weekly Wage",  
 caption="Industrial Private Sector Q4 2022",  
 x="Average employee count",  
 y="Average weekly Wage",)



dat.private.state<-data.frame(dat[dat$own\_code == 5 & dat$agglvl\_code == 55 & dat$disclosure\_code != "N" & dat$avg\_qtrly\_emplvl > 4000 & dat$avg\_wkly\_wage<2000,])  
dat.private.state$area\_title<-substr(dat.private.state$area\_title, 1, nchar(dat.private.state$area\_title) - 12)

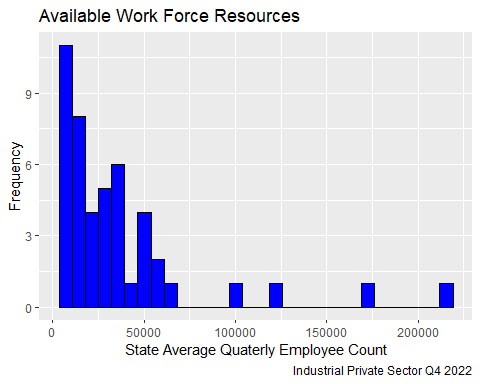
**Remove States that are not in the Continental United States**

After filtering our data set to fit the parameters that we are looking for we still have a few areas that will not be suitable for supply chain budgetary constraint for construction. These states are Hawaii, Puerto Rico, and Alaska. We will remove them from our data set as we continue our analysis.

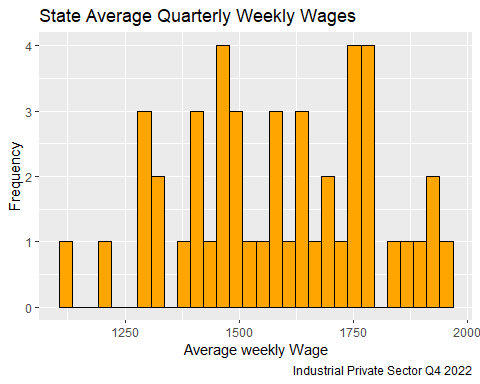
rownames(dat.private.state)<-1:nrow(dat.private.state)  
dat.private.state<-dat.private.state[-c(2,10,49),]

We will continue evaluating the frequency of occurrences of Available Work Force Resources and Average Quarterly Weekly Wages

a.state<-ggplot(data=dat.private.state,aes(x=avg\_qtrly\_emplvl))+geom\_histogram(color="black", fill="blue", bins=30)+  
 labs(  
 title="Available Work Force Resources",  
 caption="Industrial Private Sector Q4 2022",  
 x="State Average Quaterly Employee Count",  
 y="Frequency",)  
a.state



b.state<-ggplot(data=dat.private.state,aes(x=avg\_wkly\_wage))+geom\_histogram(color="black", fill="orange", bins=30)+  
 labs(   
 title="State Average Quarterly Weekly Wages",  
 caption="Industrial Private Sector Q4 2022",  
 x="Average weekly Wage",  
 y="Frequency",)  
b.state



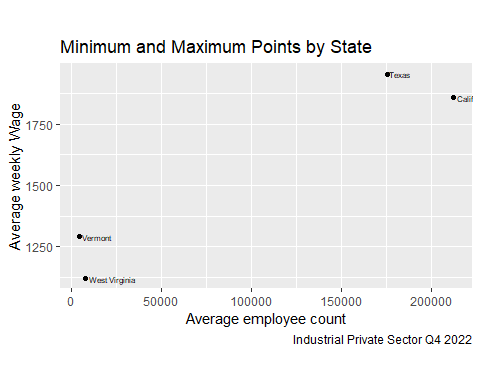
**Identify Min and Max**

As we evaluate the MIN and MAX for both variables we identify that they drive our model from a Normal distribution but have the sufficient parameters for further evaluation.The Weekly Wage that fits our model fits better into a Normal Distribution, but we can see that the average employee level clearly does not by first looking at our histograms.

weekly.wage.min <- which.min(dat.private.state$avg\_wkly\_wage)  
weekly.wage.max <- which.max(dat.private.state$avg\_wkly\_wage)  
emplvl.min <- which.min(dat.private.state$avg\_qtrly\_emplvl)  
emplvl.max <- which.max(dat.private.state$avg\_qtrly\_emplvl)  
  
ind<-c(emplvl.min, emplvl.max,weekly.wage.min, weekly.wage.max)  
dat.text<-data.frame(area\_title = dat.private.state$area\_title[ind], avg\_qtrly\_emplv = dat.private.state$avg\_qtrly\_emplv[ind], avg\_wkly\_wage = dat.private.state$avg\_wkly\_wage[ind])  
print(dat.text)

## area\_title avg\_qtrly\_emplv avg\_wkly\_wage  
## 1 Vermont 4478 1290  
## 2 California 212759 1856  
## 3 West Virginia 7747 1120  
## 4 Texas 175896 1953

g <- ggplot(dat.text, aes(x = avg\_qtrly\_emplv, y = avg\_wkly\_wage)) +  
geom\_point()+  
 geom\_text(data=dat.text,aes(x = avg\_qtrly\_emplv, y = avg\_wkly\_wage,label=area\_title),position = position\_dodge(width=4),hjust=-0.1,size=2)+  
 theme(aspect.ratio = 6/11)+labs(   
 title="Minimum and Maximum Points by State",  
 caption="Industrial Private Sector Q4 2022",  
 x="Average employee count",  
 y="Average weekly Wage",)  
   
g



high.ww<-dat.private.state[order(dat.private.state$avg\_wkly\_wage,decreasing=TRUE),][1:5,]  
(high.ww.state<-data.frame(high.ww$area\_title,high.ww$avg\_wkly\_wage))

## high.ww.area\_title high.ww.avg\_wkly\_wage  
## 1 Texas 1953  
## 2 Colorado 1920  
## 3 Tennessee 1913  
## 4 Connecticut 1890  
## 5 California 1856

low.ww<-dat.private.state[order(dat.private.state$avg\_wkly\_wage,decreasing=FALSE),][1:5,]  
(low.ww.state<-data.frame(low.ww$area\_title,low.ww$avg\_wkly\_wage))

## low.ww.area\_title low.ww.avg\_wkly\_wage  
## 1 West Virginia 1120  
## 2 South Dakota 1218  
## 3 Vermont 1290  
## 4 Maine 1295  
## 5 Idaho 1299

high.emplvl<-dat.private.state[order(dat.private.state$avg\_qtrly\_emplvl,decreasing=TRUE),][1:5,]  
(high.emplvl.state<-data.frame(high.emplvl$area\_title,high.emplvl$avg\_qtrly\_emplvl))

## high.emplvl.area\_title high.emplvl.avg\_qtrly\_emplvl  
## 1 California 212759  
## 2 Texas 175896  
## 3 Florida 125188  
## 4 New York 103718  
## 5 Pennsylvania 63598

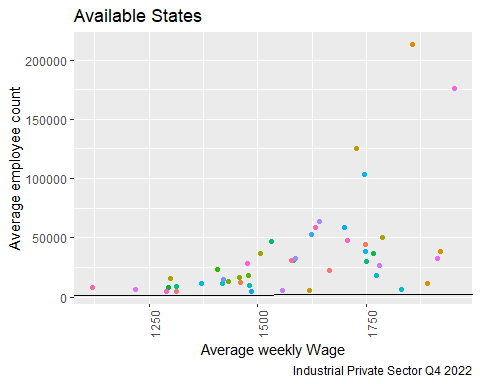
low.emplvl<-dat.private.state[order(dat.private.state$avg\_qtrly\_emplvl,decreasing=FALSE),][1:5,]  
(low.emplvl.state<-data.frame(low.emplvl$area\_title,low.emplvl$avg\_qtrly\_emplvl))

## low.emplvl.area\_title low.emplvl.avg\_qtrly\_emplvl  
## 1 Vermont 4478  
## 2 Wyoming 4545  
## 3 North Dakota 5047  
## 4 Rhode Island 5227  
## 5 Delaware 5637

**State Linearity Assessment**

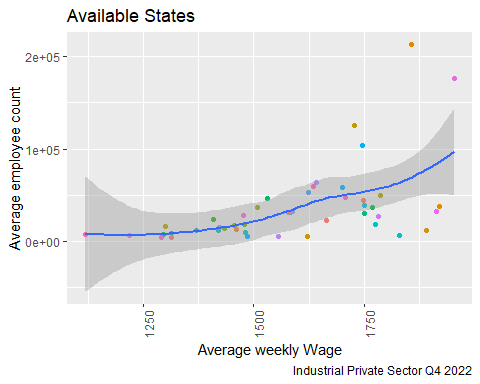
Using the geom\_abline() function we identify that our data has clustering that reduces the effectiveness of this function. For thisstraight “linear model” line is not a good fit.

ggplot(data= dat.private.state, mapping = aes(x = avg\_wkly\_wage,y = avg\_qtrly\_emplvl ))+theme(axis.text.x = element\_text(angle = 90))+geom\_point(mapping = aes(color = area\_title))+geom\_abline()+labs(   
 title="Available States",  
 caption="Industrial Private Sector Q4 2022",  
 x= "Average weekly Wage",  
 y= "Average employee count",)+ theme(legend.position="none")



We turn to geom\_smooth() function, we will use LOESS (Locally Weighted Scatterplot Smoothing) line smoothing by default since there are fewer than 1000 observations.The formula allows use to specify an exact formula to use for the smoothing line.For example, you could explicitly set “formula = y ~ x“ which we will use in this model. As we see we still have a few states that are within our models parameters, but are outside of our graphs span. This tells us that our model still has a large range even after our initial location reduction.

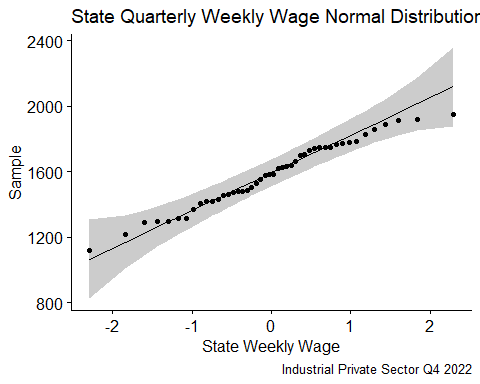
ggplot(data= dat.private.state, mapping = aes(x = avg\_wkly\_wage,y = avg\_qtrly\_emplvl ))+theme(axis.text.x = element\_text(angle = 90))+geom\_point(mapping = aes(color = area\_title))+geom\_smooth(method = loess, formula = y~x)+labs(   
 title="Available States",  
 caption="Industrial Private Sector Q4 2022",  
 x= "Average weekly Wage",  
 y= "Average employee count",)+ theme(legend.position="none")



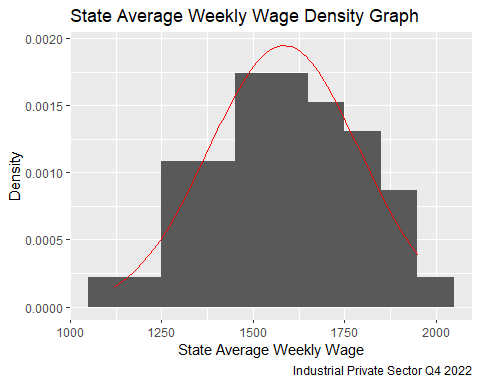
**State Normal Distribution Comparison**

As we see below, our current data for the Quarterly Employee Level per State doesn’t appear to reflect the Normal Distribution. The Average Weekly Wage Density Graph is closer to the Normal distribution than the Quarterly Employee Level Density Graph. The Quarterly Employee Level Density Graph shows the impact of higher destiny populations that may have similar wages but a larger work force.

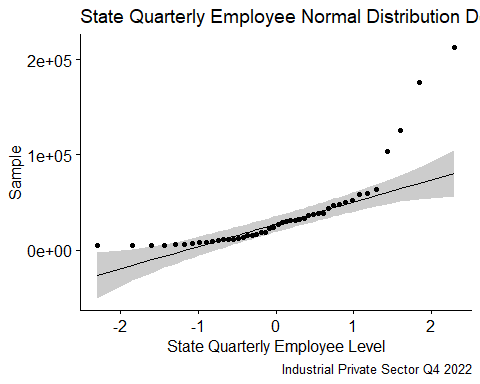
mu.ww<-mean(dat.private.state$avg\_wkly\_wage)  
sd.ww<-sqrt(sum((dat.private.state$avg\_wkly\_wage-mean(dat.private.state$avg\_wkly\_wage))^2/(length(dat.private.state$avg\_wkly\_wage)-1)))  
c<-ggqqplot(dat.private.state$avg\_wkly\_wage)+  
 labs(  
 title="State Quarterly Weekly Wage Normal Distribution Density Graph",  
 caption="Industrial Private Sector Q4 2022",  
 x="State Weekly Wage",)  
c



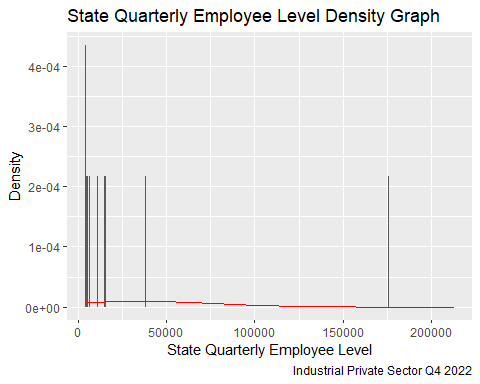
d<-ggplot(data=dat.private.state,aes(x=avg\_wkly\_wage))+  
 geom\_histogram(aes(y = after\_stat(density)),binwidth=100)+  
 stat\_function(fun = dnorm, colour = "red",args = list(mean =mu.ww, sd = sd.ww))+  
 labs(  
 title="State Average Weekly Wage Density Graph",  
 caption="Industrial Private Sector Q4 2022",  
 x="State Average Weekly Wage",  
 y="Density",)  
d



mu.el<-mean(dat.private.state$avg\_qtrly\_emplvl)  
sd.el<-sqrt(sum((dat.private.state$avg\_qtrly\_emplvl-mean(dat.private.state$avg\_qtrly\_emplvl))^2/(length(dat.private.state$avg\_qtrly\_emplvl)-1)))  
f<-ggqqplot(dat.private.state$avg\_qtrly\_emplv)+  
 labs(  
 title="State Quarterly Employee Normal Distribution Density Graph",  
 caption="Industrial Private Sector Q4 2022",  
 x="State Quarterly Employee Level",)  
f



e<-ggplot(data=dat.private.state,aes(x=avg\_qtrly\_emplvl))+  
 geom\_histogram(aes(y = after\_stat(density)),binwidth=100)+  
 stat\_function(fun = dnorm, colour = "red",args = list(mean =mu.el, sd = sd.el))+  
 labs(  
 title="State Quarterly Employee Level Density Graph",  
 caption="Industrial Private Sector Q4 2022",  
 x="State Quarterly Employee Level",  
 y="Density",)  
e



**State Count Data A Regression with Poisson Errors**

The analysis will begin with applying the regression with Poisson Errors to our state data set, using the Average Weekly Wage as our response variable.

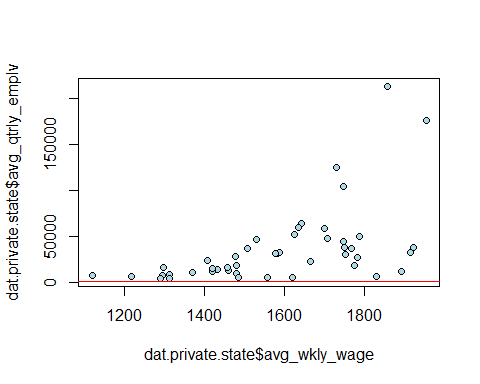
model1<-glm(dat.private.state$avg\_wkly\_wage~dat.private.state$avg\_qtrly\_emplv,poisson)  
summary(model1)

##   
## Call:  
## glm(formula = dat.private.state$avg\_wkly\_wage ~ dat.private.state$avg\_qtrly\_emplv,   
## family = poisson)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 7.312e+00 4.884e-03 1497.17 <2e-16 \*\*\*  
## dat.private.state$avg\_qtrly\_emplv 1.492e-06 8.143e-08 18.33 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 1208.53 on 45 degrees of freedom  
## Residual deviance: 890.27 on 44 degrees of freedom  
## AIC: 1317.3  
##   
## Number of Fisher Scoring iterations: 4

model1$coefficients

## (Intercept) dat.private.state$avg\_qtrly\_emplv   
## 7.311596e+00 1.492468e-06

xv<-0:2290   
yv<-7.344950+0.000001313803\*xv  
y<-exp(yv)  
plot(dat.private.state$avg\_wkly\_wage,dat.private.state$avg\_qtrly\_emplv,pch=21,bg="lightblue")  
lines(xv,y,col="red")



deviance(model1)

## [1] 890.2677

(df <- nrow(dat.private.state) - length(model1$coefficients))

## [1] 44

The residual deviance (1299.4) is much larger than the residual degrees of freedom (48). This indicates we have extra or unexplained variation in the response variable. This is referred to as overdispersion. This is also seen in our p-value, which in our case is practically 0. For this model a p-value < .05 means we have overdispersion. We can compensate for the overdispersion in our data by attempting to refit the model using quasipoisson errors.

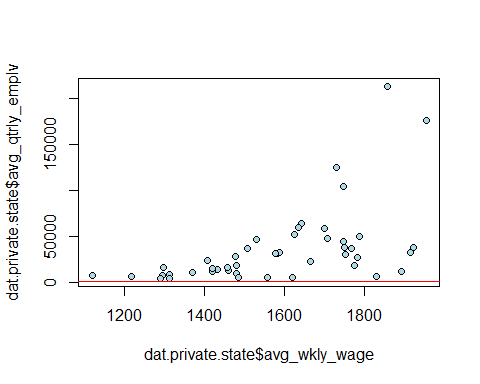
model2<-glm(dat.private.state$avg\_wkly\_wage~dat.private.state$avg\_qtrly\_emplv,quasipoisson)  
summary(model2)

##   
## Call:  
## glm(formula = dat.private.state$avg\_wkly\_wage ~ dat.private.state$avg\_qtrly\_emplv,   
## family = quasipoisson)  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.312e+00 2.199e-02 332.48 < 2e-16 \*\*\*  
## dat.private.state$avg\_qtrly\_emplv 1.492e-06 3.667e-07 4.07 0.000192 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for quasipoisson family taken to be 20.27765)  
##   
## Null deviance: 1208.53 on 45 degrees of freedom  
## Residual deviance: 890.27 on 44 degrees of freedom  
## AIC: NA  
##   
## Number of Fisher Scoring iterations: 4

model2$coefficients

## (Intercept) dat.private.state$avg\_qtrly\_emplv   
## 7.311596e+00 1.492468e-06

xv<-0:2290   
yv<-7.344950+0.000001313803\*xv  
y<-exp(yv)  
plot(dat.private.state$avg\_wkly\_wage,dat.private.state$avg\_qtrly\_emplv,pch=21,bg="lightblue")  
lines(xv,y,col="red")



deviance(model2)

## [1] 890.2677

(df <- nrow(dat.private.state) - length(model2$coefficients))

## [1] 44

The residual deviance (1299.4) and the residual degrees of freedom (48) did not change with the new model. This indicates we still have overdispersion in our data. Even though the p value increased (p = .0012), it is still less than the desired value of .05. Thus our model is still not a good fit for our data.

one.way<- aov(avg\_wkly\_wage~avg\_qtrly\_emplvl, data = dat.private.state)  
  
summary.lm(one.way)

##   
## Call:  
## aov(formula = avg\_wkly\_wage ~ avg\_qtrly\_emplvl, data = dat.private.state)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -390.85 -98.94 -12.41 116.23 369.25   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.491e+03 3.423e+01 43.556 < 2e-16 \*\*\*  
## avg\_qtrly\_emplvl 2.570e-03 6.179e-04 4.158 0.000146 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 175.6 on 44 degrees of freedom  
## Multiple R-squared: 0.2821, Adjusted R-squared: 0.2658   
## F-statistic: 17.29 on 1 and 44 DF, p-value: 0.0001459

qf(0.95, 1, 48)

## [1] 4.042652

shapiro.test(one.way$residuals) #will probably not need

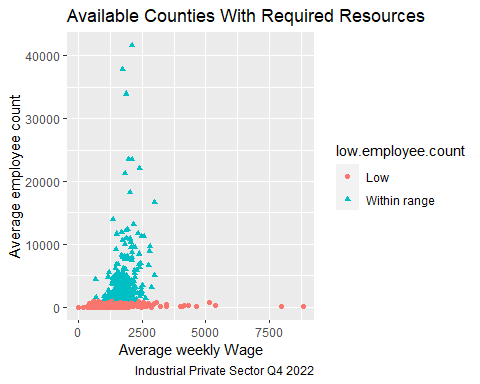
##   
## Shapiro-Wilk normality test  
##   
## data: one.way$residuals  
## W = 0.97829, p-value = 0.5377

The calculated value of the F test statistic of 12.22 is greater than the critical value of F = 4.04. Thus we can reject our null hypothesis and conclude that wage is affected by employee level.

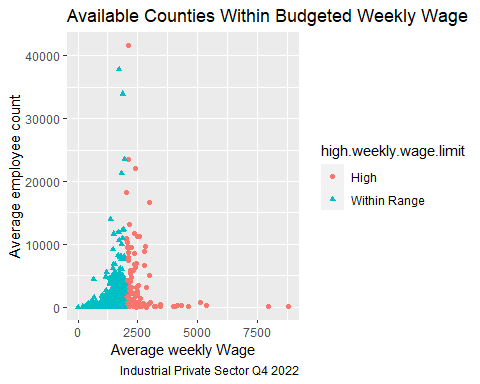
The Shapiro-Wilk normality test was done to verify whether or not our residuals come from a normal distribution. Since our p-value for this test is greater than .05, we do not reject the null hypothesis, and thus the residual data do follow a normal distribution.

# County Available Workforce by Weekly Wage Analysis

dat.private.county<-data.frame(dat[dat$own\_code == 5 & dat$agglvl\_code == 75 & dat$disclosure\_code != "N",])  
dat.private.county$low.employee.count<-ifelse(dat.private.county$avg\_qtrly\_emplvl>=1000,"Within range", "Low")  
dat.private.county$high.weekly.wage.limit<-ifelse(dat.private.county$avg\_wkly\_wage<2000,"Within Range","High")  
  
  
ggplot(data= dat.private.county, mapping = aes(x = avg\_wkly\_wage,y = avg\_qtrly\_emplvl ))+geom\_point(mapping = aes(color = low.employee.count, shape = low.employee.count ))+labs(   
 title="Available Counties With Required Resources",  
 caption="Industrial Private Sector Q4 2022",  
 x="Average weekly Wage",  
 y="Average employee count",)



ggplot(data= dat.private.county, mapping = aes(x = avg\_wkly\_wage,y = avg\_qtrly\_emplvl ))+geom\_point(mapping = aes(color = high.weekly.wage.limit, shape = high.weekly.wage.limit ))+labs(   
 title="Available Counties Within Budgeted Weekly Wage",  
 caption="Industrial Private Sector Q4 2022",  
 x="Average weekly Wage",  
 y="Average employee count",)



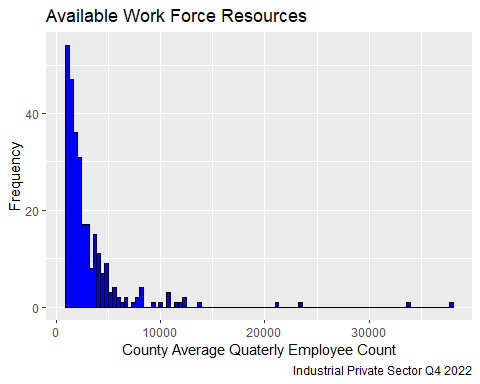
**Remove Counties that are not in the Continental United States**

As we did with the state analysis we are going to remove the counties located in Hawaii, Puerto Rico, and Alaska. We will remove them from our data set as we continue our analysis. And build our data set to include the areas that fit our model.

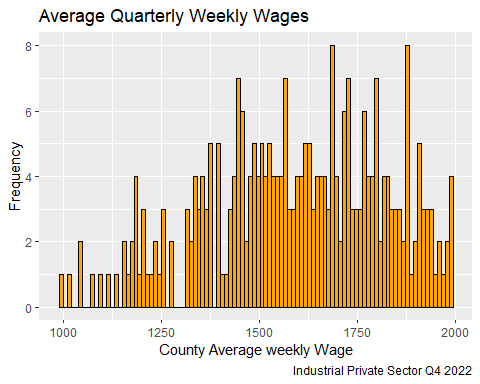
dat.private.county<-data.frame(dat[dat$own\_code == 5 & dat$agglvl\_code == 75 & dat$disclosure\_code != "N"& dat$avg\_qtrly\_emplvl > 1000 & dat$avg\_wkly\_wage<2000,])  
dat.private.county<-dat.private.county[!grepl("Alaska", dat.private.county$area\_title),]  
dat.private.county<-dat.private.county[!grepl("Puerto", dat.private.county$area\_title),]  
dat.private.county<-dat.private.county[!grepl("Hawaii", dat.private.county$area\_title),]

We now wish to analyze the data by county instead of by state and see if our model changes based on a more specific location.

a.county<-ggplot(data=dat.private.county,aes(x=avg\_qtrly\_emplvl))+geom\_histogram(color="black", fill="blue", bins=100)+  
 labs(  
 title="Available Work Force Resources",  
 caption="Industrial Private Sector Q4 2022",  
 x="County Average Quaterly Employee Count",  
 y="Frequency",)  
a.county



b.county<-ggplot(data=dat.private.county,aes(x=avg\_wkly\_wage))+geom\_histogram(color="black", fill="orange", bins=100)+  
 labs(   
 title="Average Quarterly Weekly Wages",  
 caption="Industrial Private Sector Q4 2022",  
 x="County Average weekly Wage",  
 y="Frequency",)  
b.county



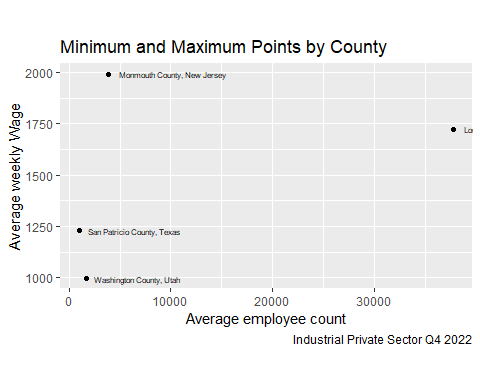
**Identify County Min and Max**

As we evaluate the MIN and MAX for both variables we identify that they drive our model from a Normal distribution but have the sufficient parameters for further evaluation.The Weekly Wage that fits our model fits better into a Normal Distribution, but we can see that the average employee level clearly does not by first looking at our histograms.

weekly.wage.min <- which.min(dat.private.county$avg\_wkly\_wage)  
weekly.wage.max <- which.max(dat.private.county$avg\_wkly\_wage)  
emplvl.min <- which.min(dat.private.county$avg\_qtrly\_emplvl)  
emplvl.max <- which.max(dat.private.county$avg\_qtrly\_emplvl)  
  
ind<-c(emplvl.min, emplvl.max,weekly.wage.min, weekly.wage.max)  
dat.text<-data.frame(area\_title = dat.private.county$area\_title[ind], avg\_qtrly\_emplv = dat.private.county$avg\_qtrly\_emplv[ind], avg\_wkly\_wage = dat.private.county$avg\_wkly\_wage[ind])  
print(dat.text)

## area\_title avg\_qtrly\_emplv avg\_wkly\_wage  
## 1 San Patricio County, Texas 1004 1229  
## 2 Los Angeles County, California 37772 1723  
## 3 Washington County, Utah 1681 997  
## 4 Monmouth County, New Jersey 3864 1991

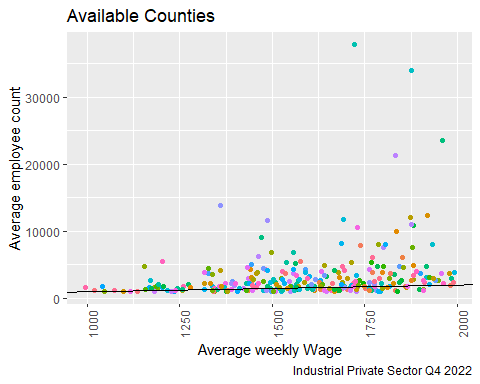
g <- ggplot(dat.text, aes(x = avg\_qtrly\_emplv, y = avg\_wkly\_wage)) +  
geom\_point()+  
 geom\_text(data=dat.text,aes(x = avg\_qtrly\_emplv, y = avg\_wkly\_wage,label=area\_title),position = position\_dodge(width=4),hjust=-0.1,size=2)+  
 theme(aspect.ratio = 6/11)+labs(   
 title="Minimum and Maximum Points by County",  
 caption="Industrial Private Sector Q4 2022",  
 x="Average employee count",  
 y="Average weekly Wage",)  
g



**County Linearity Assessment**

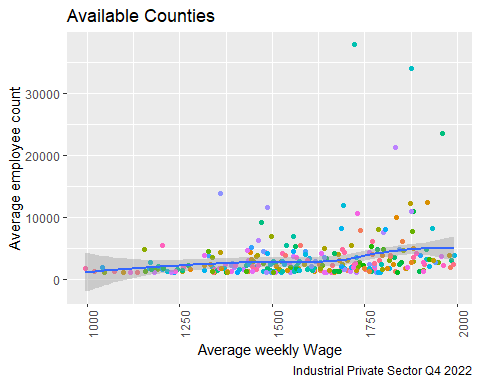
Using the geom\_abline() function we identify that our data has clustering that reduces the effectiveness of this function. For this straight “linear model” line is a better fit to the county analysis in comparison to the state analysis.

ggplot(data= dat.private.county, mapping = aes(x = avg\_wkly\_wage,y = avg\_qtrly\_emplvl ))+theme(axis.text.x = element\_text(angle = 90))+geom\_point(mapping = aes(color = area\_title))+geom\_abline()+labs(   
 title="Available Counties",  
 caption="Industrial Private Sector Q4 2022",  
 x= "Average weekly Wage",  
 y= "Average employee count",)+ theme(legend.position="none")



Once again we turn to geom\_smooth() function, we will use LOESS (Locally Weighted Scatterplot Smoothing) line smoothing by default since there are fewer than 1000 observations.The formula allows use to specify an exact formula to use for the smoothing line.For example, you could explicitly set “formula = y ~ x“ which we will use in this model. As we see we have a greater amount of counties that are within our models parameters, but there are still a few outside of our graphs span as well. The county model looks promising in comparison the the higher level we see in the state analysis.

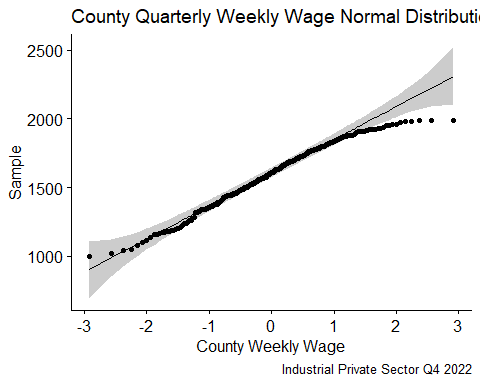
ggplot(data= dat.private.county, mapping = aes(x = avg\_wkly\_wage,y = avg\_qtrly\_emplvl ),)+theme(axis.text.x = element\_text(angle = 90))+geom\_point(mapping = aes(color = area\_title))+geom\_smooth(method = loess, formula = y~x)+labs(   
 title="Available Counties",  
 caption="Industrial Private Sector Q4 2022",  
 x= "Average weekly Wage",  
 y= "Average employee count",)+theme(legend.position="none")



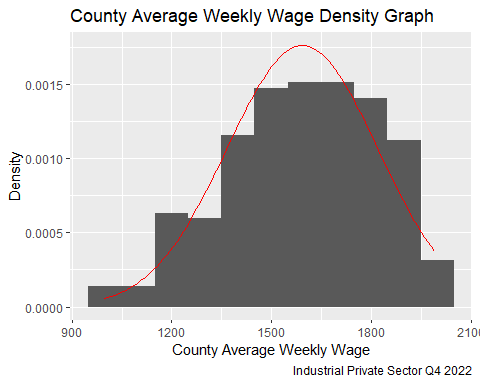
**County Normal Distribution Comparison**

As we see below, our current data for the Quarterly Employee Level per county doesn’t appear to reflect the Normal Distribution. The Average Weekly Wage Density Graph is closer to the Normal distribution than the Quarterly Employee Level Density Graph. The Quarterly Employee Level Density Graph shows the impact of higher destiny populations that may have similar wages but a larger work force. We see the same result as before during the state level analysis.

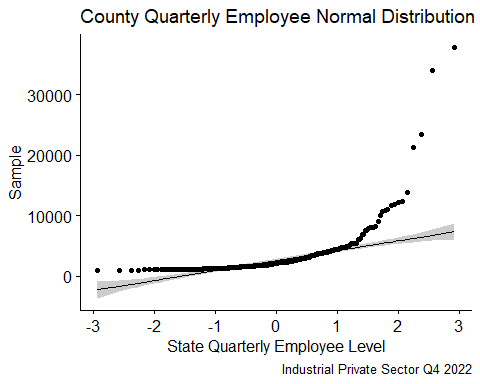
mu.ww<-mean(dat.private.county$avg\_wkly\_wage)  
sd.ww<-sqrt(sum((dat.private.county$avg\_wkly\_wage-mean(dat.private.county$avg\_wkly\_wage))^2/(length(dat.private.county$avg\_wkly\_wage)-1)))  
c<-ggqqplot(dat.private.county$avg\_wkly\_wage)+  
 labs(  
 title="County Quarterly Weekly Wage Normal Distribution Density Graph",  
 caption="Industrial Private Sector Q4 2022",  
 x="County Weekly Wage",)  
c



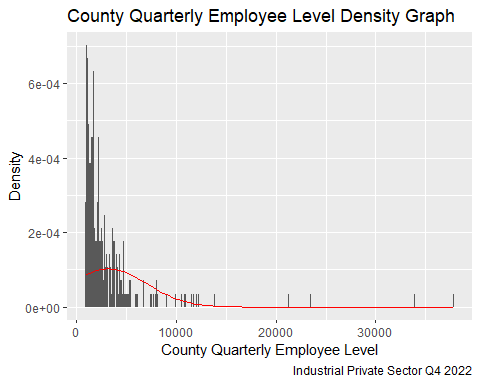
d<-ggplot(data=dat.private.county,aes(x=avg\_wkly\_wage))+  
 geom\_histogram(aes(y = after\_stat(density)),binwidth=100)+  
 stat\_function(fun = dnorm, colour = "red",args = list(mean =mu.ww, sd = sd.ww))+  
 labs(  
 title="County Average Weekly Wage Density Graph",  
 caption="Industrial Private Sector Q4 2022",  
 x="County Average Weekly Wage",  
 y="Density",)  
d



mu.el<-mean(dat.private.county$avg\_qtrly\_emplvl)  
sd.el<-sqrt(sum((dat.private.county$avg\_qtrly\_emplvl-mean(dat.private.county$avg\_qtrly\_emplvl))^2/(length(dat.private.county$avg\_qtrly\_emplvl)-1)))  
f<-ggqqplot(dat.private.county$avg\_qtrly\_emplv)+  
 labs(  
 title="County Quarterly Employee Normal Distribution Density Graph",  
 caption="Industrial Private Sector Q4 2022",  
 x="State Quarterly Employee Level",)  
f



e<-ggplot(data=dat.private.county,aes(x=avg\_qtrly\_emplvl))+  
 geom\_histogram(aes(y = after\_stat(density)),binwidth=100)+  
 stat\_function(fun = dnorm, colour = "red",args = list(mean =mu.el, sd = sd.el))+  
 labs(  
 title="County Quarterly Employee Level Density Graph",  
 caption="Industrial Private Sector Q4 2022",  
 x="County Quarterly Employee Level",  
 y="Density",)  
e



**County Count Data A regression with Poisson Errors**

As with the State analysis, Average Weekly Wage is our response variable.

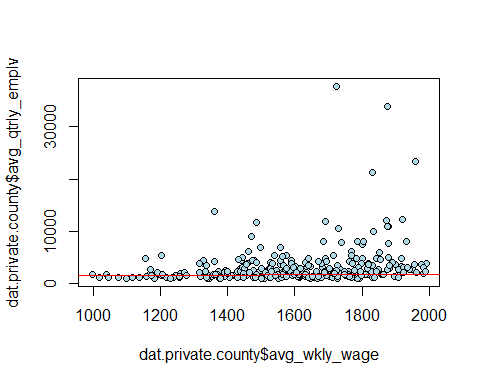
model3<-glm(dat.private.county$avg\_wkly\_wage~dat.private.county$avg\_qtrly\_emplv,poisson)  
summary(model3)

##   
## Call:  
## glm(formula = dat.private.county$avg\_wkly\_wage ~ dat.private.county$avg\_qtrly\_emplv,   
## family = poisson)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 7.347e+00 1.893e-03 3881.18 <2e-16 \*\*\*  
## dat.private.county$avg\_qtrly\_emplv 8.033e-06 3.500e-07 22.95 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 9405.0 on 284 degrees of freedom  
## Residual deviance: 8907.6 on 283 degrees of freedom  
## AIC: 11534  
##   
## Number of Fisher Scoring iterations: 4

model3$coefficients

## (Intercept) dat.private.county$avg\_qtrly\_emplv   
## 7.346851e+00 8.033384e-06

xv<-0:2994   
yv<-7.392957+0.00001246112\*xv  
y<-exp(yv)  
plot(dat.private.county$avg\_wkly\_wage,dat.private.county$avg\_qtrly\_emplv,pch=21,bg="lightblue")  
lines(xv,y,col="red")



deviance(model3)

## [1] 8907.586

(df <- nrow(dat.private.county) - length(model3$coefficients))

## [1] 283

The residual deviance (23111) is much larger than the residual degrees of freedom (351). This indicates we have overdispersion in this model as well. Once again confirmed by a p-value < .05. We can compensate for the overdispersion in our data by attempting to refit the model using quasipoisson errors.

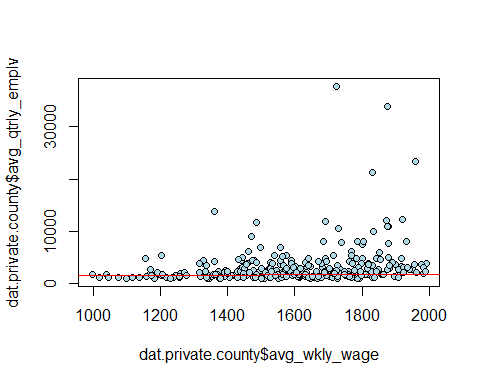
model4<-glm(dat.private.county$avg\_wkly\_wage~dat.private.county$avg\_qtrly\_emplv,quasipoisson)  
summary(model4)

##   
## Call:  
## glm(formula = dat.private.county$avg\_wkly\_wage ~ dat.private.county$avg\_qtrly\_emplv,   
## family = quasipoisson)  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.347e+00 1.050e-02 699.866 < 2e-16 \*\*\*  
## dat.private.county$avg\_qtrly\_emplv 8.033e-06 1.941e-06 4.139 4.61e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for quasipoisson family taken to be 30.75371)  
##   
## Null deviance: 9405.0 on 284 degrees of freedom  
## Residual deviance: 8907.6 on 283 degrees of freedom  
## AIC: NA  
##   
## Number of Fisher Scoring iterations: 4

model4$coefficients

## (Intercept) dat.private.county$avg\_qtrly\_emplv   
## 7.346851e+00 8.033384e-06

xv<-0:2994   
yv<-7.392957+0.00001246112\*xv  
y<-exp(yv)  
plot(dat.private.county$avg\_wkly\_wage,dat.private.county$avg\_qtrly\_emplv,pch=21,bg="lightblue")  
lines(xv,y,col="red")



deviance(model4)

## [1] 8907.586

(df <- nrow(dat.private.county) - length(model4$coefficients))

## [1] 283

The residual deviance (23111) and the residual degrees of freedom (351) did not change with the new model. This indicates we still have overdispersion in our data. While the p-value did increase, it is still much smaller than .05. Therefore, our model at the county level is also probably not a good fit for our data.

one.way3<- aov(avg\_wkly\_wage~avg\_qtrly\_emplvl, data = dat.private.county)  
  
summary.lm(one.way3)

##   
## Call:  
## aov(formula = avg\_wkly\_wage ~ avg\_qtrly\_emplvl, data = dat.private.county)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -574.30 -148.04 24.36 167.30 407.40   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.548e+03 1.697e+01 91.228 < 2e-16 \*\*\*  
## avg\_qtrly\_emplvl 1.395e-02 3.358e-03 4.156 4.3e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 220.5 on 283 degrees of freedom  
## Multiple R-squared: 0.05751, Adjusted R-squared: 0.05418   
## F-statistic: 17.27 on 1 and 283 DF, p-value: 4.302e-05

qf(0.95, 1, 48)

## [1] 4.042652

The calculated value of the F test statistic of 40.58 is greater than the critical value of F = 4.04. Thus we can reject our null hypothesis and conclude that employee level is affected by average weekly wage.

# Conclusion

**Null hypothesis:**

The average quarterly weekly wage is not affected by the size of the employee level availability by location.

location makes no difference

**Alt hypothesis:**

The average quarterly weekly wage is affected by the size of employee level availability by location.

location does make a difference

**Our research question was:**

How does the weekly wage affect the cost of building a single design industrial complex for a budgeted construction contract with a suitable workforce located in the United States?

**Process Results**

We performed poisson regression on our data at the state and county levels to check for a correlation and then ran the one way ANOVA test to further dive into our data.

When analyzing our data at the state level, there was enough evidence to support rejecting the null hypothesis.

When analyzing our data at the county level, there was enough evidence to support rejecting the null hypothesis.

Taking into consideration the over dispersion of our data at both levels, it is feasible that there is an effect from the size of the employee level availability to the average weekly wage but it appears as if there are likely other factors involved as well.

This seems to be a good starting point when answering our research question, however, in reality we would need to look into other factors when making this decision.

**Additional Factors for consideration:**

The data set references the NAICS 236 Construction of buildings but does not separate by skilled(trade specialties), unskilled, or if wages represented Union labor wage considerations.

Additional location analysis for cost of purchasing property in those location would also have to factored in to identify if location is within budgeted constraints.

Suitable location considerations also need to be considered for hiring personnel to operate the plant along with additional vendor support for equipment maintenance. If we would build a plant in the middle of nowhere would that support our needs or do we need to build close to a large metropolitan area for supplies and support requirements?

**Next Steps**

This analysis allows for the ability to capture the mean of wage considerations when planning a budget and long term construction plan. It identifies the areas for consideration to begin identifying suitable locations for developing a widespread infrastructure.

With additional analysis of all factors stated the infrastructure can be developed. Since construction and economical factors shift we must return to this analysis when new data is available to ensure that budget forecasting models are sustainable or implement design changes that may be required.