Homework 4

STA 309

1 – Peak Power Consumption

QUESTION: Build a linear model for peak power that includes the main effects (temp, weekday) and the interaction, to answer the following:

I. During the summer, how much higher or lower is daytime power consumption on a weekday versus a weekend?

II. During the summer, how does daytime power consumption increase with temperature, on average, and does this relationship seem to differ between weekends and weekdays?

III. Provide a faceted scatter plot that shows the relationship between power consumption and temperature, faceted by weekday.

APPROACH: To answer this question, I fitted a linear model, while bootstrapping, over a Monte Carlo simulation, regressing the model 10,000 times. The variables for the linear model were temperate, the weekday dummy variable, and the interaction between the two. Then determining the confidence intervals for this approach.

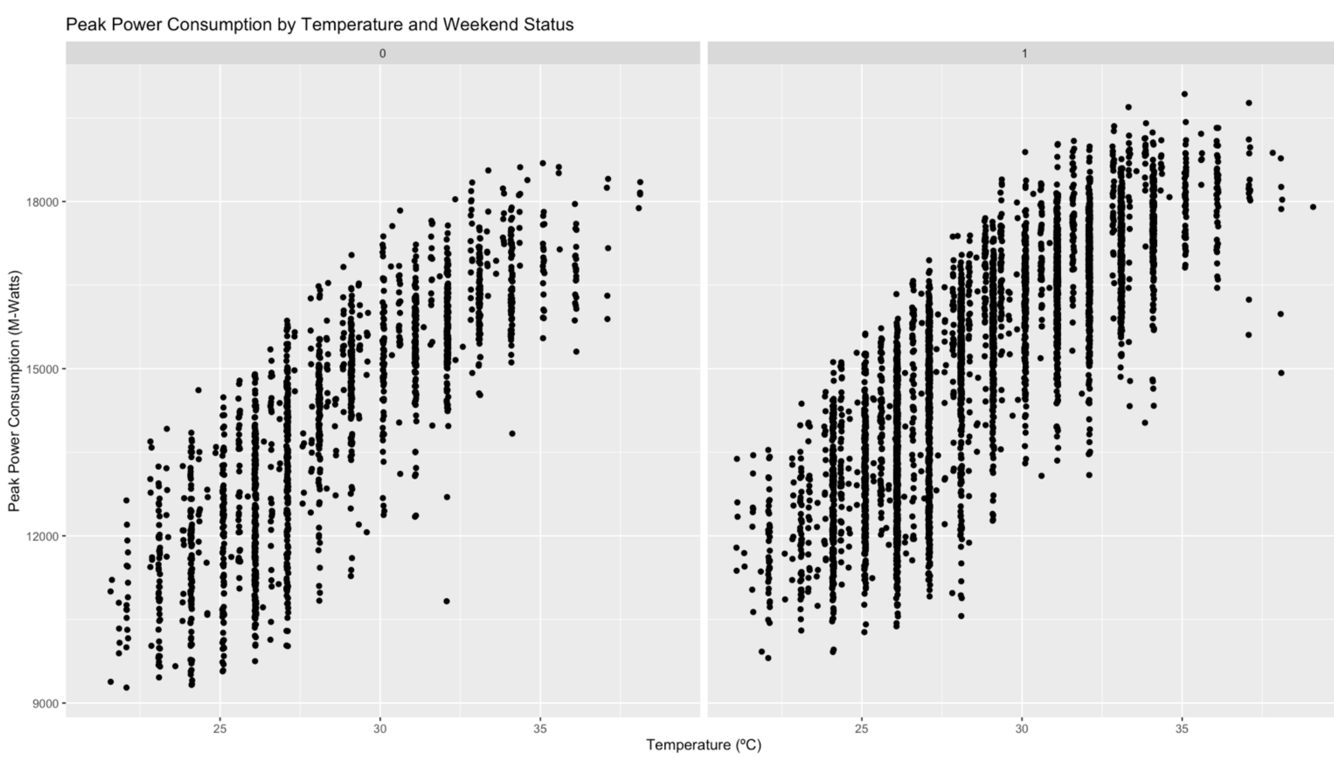
RESULTS:

FIGURE 1

-	Lower	Upper	Level	Estimate
Intercept	-1035.119	-120.778	95%	-569.919
Temperature	496.989	527.827	95%	512.214
Weekday	422.925	1507.177	95%	962.098
Temperature:Weekday	-18.080	18.717	95%	0.461

I. This table shows the confidence intervals for each of the variables in the linear regression model, which was bootstrapped, and their respective estimates.

FIGURE 2



I. The scatter plot above shows the peak power consumption in m-watts on the y-axis and the temperature in ºC on the x-axis. This plot is faceted by weekday status, with 0 being the weekend and 1 being the weekday.

CONCLUSION: In light of the results presented, it can be determined that at a 95% confidence level that power usage would increase anywhere between 496.99 to 527.83, estimated to be 512.21 per degree increase in temperature; the power usage would increase anywhere between 422.95 and 1507.18, estimated to be 962.10 on a weekday; when it is a weekday, a temperature increase is between -18.08 and 18.717, estimated to be 0.461. The interaction appears to be inconclusive as the confidence interval is from -18 to 18, causing it to be a very small difference in predictability.

I. With 95% confidence, power consumption on a weekday is estimated to be, determined by the confidence intervals previously stated, between 496.99 and 528.24 (original + interaction) m-Watts higher; this is when you add the interaction to the estimated offset of the dummy variable.
II. With 95% confidence, power consumption increases between 422.93 and 1507.63 (original + interaction), on average 512.214 (or 512.67 when taking into account the interaction between it being a weekday), when taking into account the interaction of the weekday variable. This differs from weekdays and weekends in that the weekday, on average, by a 422.925 to 1507.18 (962.10 estimate) increase in power usage when it is a weekday, or 422.95 to 1507.63 (962.55) increase when the interaction between temperature and it being a weekday is accounted for.

2 – UT Real Estate

QUESTION: Build a linear regression model for the rent of a two-bedroom apartment in West Campus in terms of the variables.

I. What is the drop-off in price associated with an extra 0.1-mile walking distance to the UT Tower, adjusting for other features of the apartment (including size of unit)?

II. How does the answer to (I) differ from an unadjusted analysis, simply regressing to distance?

III. What is the increase in rent associated with an additional 100 sqft in size, adjusting for other features?

IV. How does the answer to (III) differ from an unadjusted analysis, simply regressing to sqft?

V. Include two scatter plots, one for rent versus distance and one for rent versus square footage

APPROACH: To answer these questions I regressed a linear model to the data, using all of the variables, bootstrapping over a Monte Carlo simulation, regressing 10,000 times. For the individualized models, I did the same method but only with one predictor variable. Then I determined the confidence intervals for the respective simulations.

RESULTS:

FIGURE 1

-	Lower	Upper	Level	Estimate
Intercept	493.92	1122.97	95%	794.77
Sqft	0.06	0.70	95%	0.38
Distance	-587.11	-164.63	95%	-379.21
Furnished	80.52	233.82	95%	163.29
Pool	-121.44	18.78	95%	-48.39
Laundry	42.16	180.91	95%	103.84
Electricity	-28.63	148.73	95%	63.64
Water	-134.59	83.70	95%	-23.25

I. This table shows the respective upper and lower intervals of each variable, including the intercept value, at a 95% confidence interval. This linear model regresses for every variable included in the data set. Dummy variables included: furnished, pool, laundry, electricity, and water.

FIGURE 2

-	Lower	Upper	Level	Estimate
Intercept	1226.17	1508.35	95%	1366.81
Distance	-842.99	-422.27	95%	-631.51

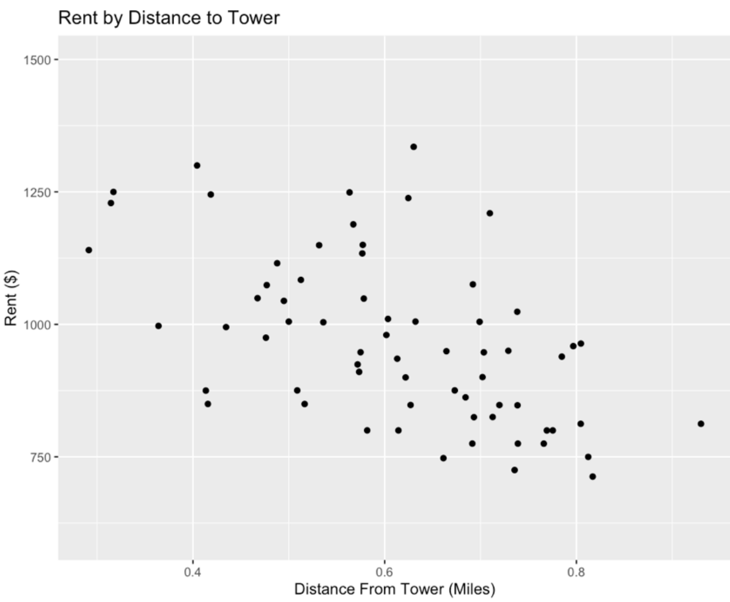
I. This table shows the respective upper and lower intervals for the Distance variable (per mile) that the price will decrease for two-bedroom rent. With the estimate being shown. All of these are at a 95% confidence.

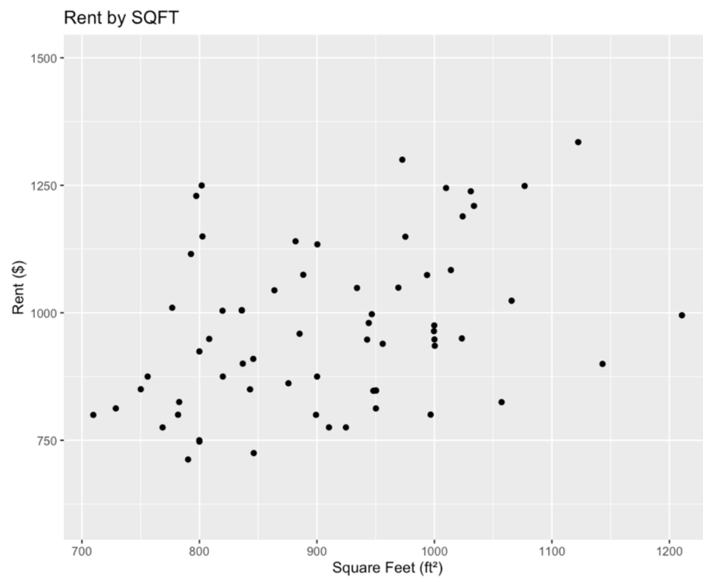
FIGURE 3

-	Lower	Upper	Level	Estimate
Intercept	179.15	838.36	0.95	518.60
SQFT	0.166	0.89	95%	0.51

I. This table shows the respective upper and lower intervals for the upper and sqft variable, meaning that per square foot increase, the price increases by that amount. This is at a 95% confidence.

FIGURE 4, 5





I. Figure 4, on the left, shows the rent on the y-axis and the distance from the tower (in miles) on the x-axis. This scatter plot attempts to establish a relationship between distance and rent.

II. Figure 5, on the right, shows the rent on the y-axis and the square feet of an apartment on the x-axis. This scatter plot attempts to establish a relationship between size and rent.

CONCLUSION: At a 95% confidence, the rent of a two-bedroom apartment in West Campus would increase between $0.06 to $0.70 per square foot increase; decrease between $587.11 and $164.63 per mile distance from the tower; increase between $80.52 and $233.82 if furnished;

increase between ($121.44) and 18.78 if the complex has a pool; increase between $42.16 and $180.91 if the laundry is in unit; increase between ($28.63) and $148.73 if electricity is included in monthly rent; and, increase between $(134.59) and $83.70 if the water is included in the rent.

I. The drop-off in price associated with an extra 0.1-mile walking distance from the UT Tower, when adjusting for other features of the apartment, is between a $16.46 to $58.71 decrease in price (or -$58.71 to -$16.71 increase); on average this drop-off in price is $37.92, per 0.1-miles from the tower.

II. If simply regressing by distance, the confidence intervals change from the previous (-$587.11, -$164.63) to (-$842.99, -$422.27). This changes the answer in question (I) from it being a $16.46 to $58.71 decrease in rent per 0.1-mile increase in distance from the tower, to a $84.30 to $42.22 decrease in rent per 0.1-mile increase in distance from the tower, averaging out to be an $63.15 decrease in rent per 0.1-mile increase.

III. The increase in rent associated with an additional 100 square feet in size, adjusting for other features of the apartment, is between a $5.94 to $69.70 increase per 100 additional square foot, averaging out to be a $37.55 increase per 100 square foot increase.

IV. If simply regressing by square-feet, the confidence intervals changes from the previous ($5.94, $69.70) to the new ($16.60, $88.58). This changes the answer in question (III) from being a $5.94 to $69.70 increase per 100 additional square foot, to a $16.60 to $88.58 increase per 100 square foot, averaging out to be a $50.96 increase per 100 square feet increase.

3 – Redlining

QUESTION: Use a linear regression model to assess whether there is an association between the number of F.A.I.R. policies and the racial/ethnic composition of a Z.I.P. code.

APPROACH: To determine whether or not there is a link between the number of policies and the racial composition of an area, I regressed three models using bootstrapping, over a Monte-Carlo simulation of 10,000 simulations. The first model accounted for all the variables excluding race, the second model accounted for all the variables including race, the third model only accounted for race. I additionally plotted all of the linear models on scatter plots.

RESULTS:

FIGURE 1

ZIP	Minority	Fire	Age	Policies	Income	No Race	With Race	Only Race
60653	99.7	21.6	65	0.9	5.583	1.442	1.280	1.465
60621	98.9	17.4	68.6	2.2	7.52	1.154	1.112	1.454
60624	94.4	18.4	72.9	1.8	7.342	1.211	1.152	1.395
60640	22.2	9.5	76.5	0.1	9.323	0.823	0.489	0.440
60657	17.3	7.7	66.9	0.5	10.656	0.600	0.336	0.376

This table shows the respective zip codes, three with a high minority percentage, and two with a low minority percentage (that did not have 0.0 in the policies section), with the respective variables – minority, fire, age, policies, income, no race, with race, and only race. No Race indicates the predicted policies per 100 units, regressed by everything except race; With Race indicates the predicted policies per 100 units, regressed by everything including race; Only Race indicates the predicted policies per 100 units, regressed by only race.

FIGURE 2

	Lower	Upper	Estimate
Intercept	0.193	2.532	0.835
Income	-0.197	-0.035	-0.075
Age	-0.003	0.008	0.003
Fire	0.006	0.066	0.031

This table shows the confidence intervals and the predicted values for the regressed values, this is for the linear regression model that did not account for race.

FIGURE 3

	Lower	Upper	Estimate
Intercept	-1.393	0.611	-0.170
Minority	0.002	0.015	0.008
Income	-0.062	0.070	-0.012
Age	0.000	0.012	0.006
Fire	0.002	0.058	0.023

This table shows the confidence intervals and the predicted values for the regressed values, this is for the linear regression model that accounted for race among the other variables.

FIGURE 4

	Lower	Upper	Estimate
Intercept	-301.552	681.763	215.499
ZIP	-0.011	0.005	-0.004
Income	-0.201	-0.032	-0.076
Age	-0.004	0.008	0.003
Fire	0.004	0.063	0.029

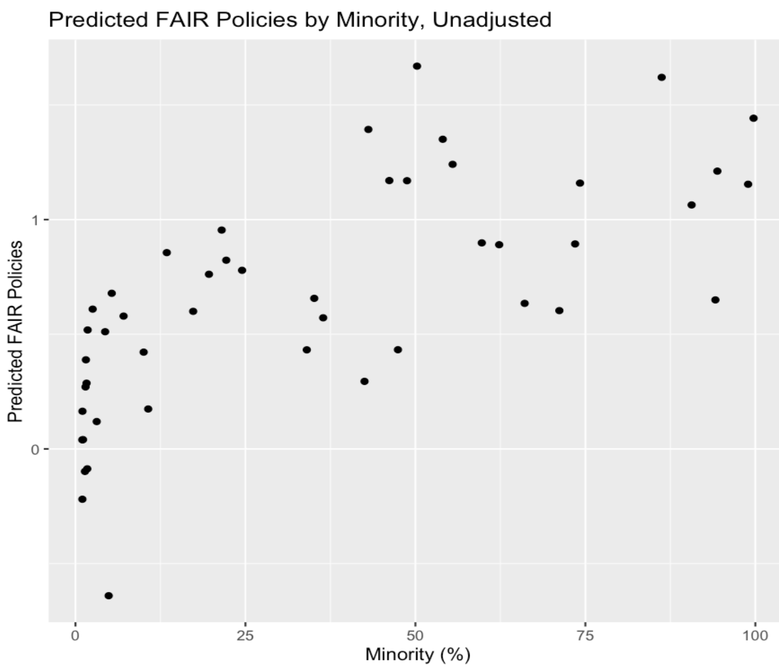
This table shows the confidence intervals and the predicted values for the regressed values, this regresses for ZIP code, to show any changes to the other predictor variables

FIGURE 5

	Lower	Upper	Estimate
Intercept	0.012	0.272	0.129
Minority	0.010	0.018	0.014

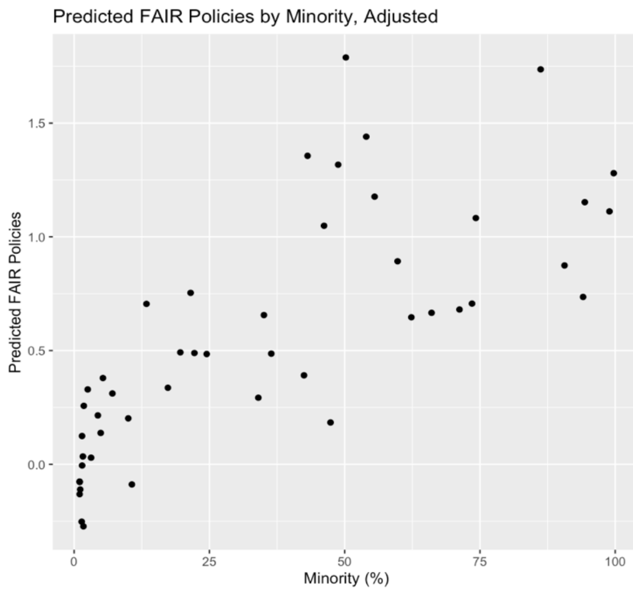
This table shows the confidence intervals and the predicted values for the regressed values, this is for the linear regression model that only accounted for race.

FIGURE 6



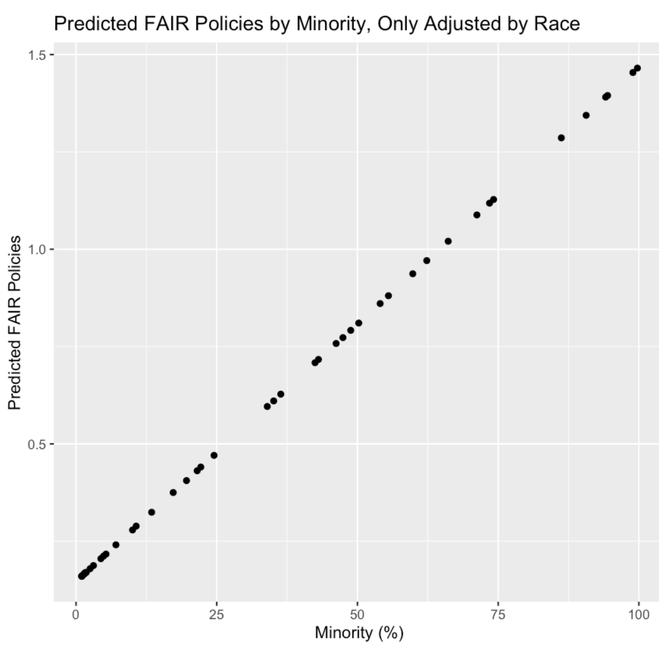
This scatter plot shows the predicted policies, by the percentage of minorities, using the linear model that did not account for race.

FIGURE 7



This scatter plot shows the predicted policies, by the percentage of minorities, using the linear model that did account for race.

FIGURE 8



This scatter plot shows the predicted policies, by the percentage of minorities, using the linear model that only accounted for race.

CONCLUSION: In light of the results presented, I would say that there **is an association** between the number of FAIR policies and racial/ethnic composition of a ZIP code. In Figure 1, the link is shown as in two of the high minority composition ZIP codes, the predicted model without accounting for race was much lower than actual; with the models accounting for race being more accurate; this shows that race is a viable predictor for the number of policies as established under Figure 1. In Figure 2, 3, 4, and 5 the confidence intervals for each of the predictor variables change when accounting for race, with the confidence being 95%. For example, the income confidence intervals for Figure 3 was (-0.197, -0.035), when accounting for race, averaging out to be -0.075 – this is, for every dollar in income, the number of FAIR policies decrease by 0.075 per 100 units, providing more access; however, when regressing for race, the income confidence intervals changes to (-0.062, 0.070), averaging out to be -0.012 – that is, for every dollar in income, the number of FAIR policies decrease by 0.012, which is quite a substantial difference from the model that did not account for race. If there was not a substantial link between the minority composition of a ZIP code and the number of FAIR policies, there should not have been such a substantial difference in the regressed values; this is shown in Figure 4, where when the model was regressed by ZIP, the estimated values to income, age, and fire were very little, if at all, changed. Finally, the scatterplots show a substantial drop in the number of predicted FAIR policies from the model that was not regressed (Figure 6) when the minority percentage was low, to the model that was regressed for race (Figure 7). Moreover, if race was not a substantial confounder, the predicted FAIR policies would not change by this amount; additionally, the linear model regressed for only race shows a positive slope (i.e. the more the race, the more the policies, in Figure 8).