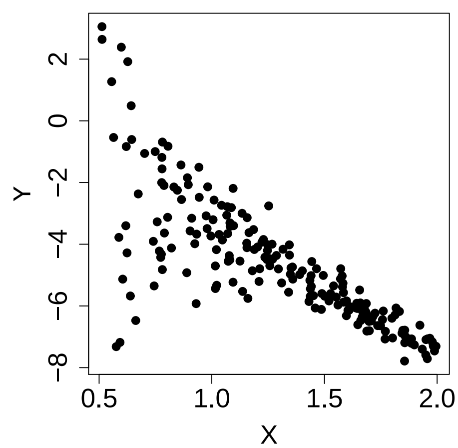
Chicago Booth BUS 41100
Solutions to Midterm Exam **SAMPLE #2**

Instructor: Max H. Farrell

These solutions are a guide only! Your answers should show more work/detail/reasoning.

1 Short Answer/Multiple Choice

(a) Use the space to the right of the plot to list which assumptions required by linear regression, if any, appear to be violated in the data set plotted below.



Solution. *The variance is not constant, and appears to decrease with X. You weren't asked to propose a solution, but here you could try dividing through by $1/X$ or $1/X^2$ (i.e. multiplying by X or X^2).*

(b) If $n = 25$, $\bar{Y} = -6$, $\bar{X} = 4$, $s_Y^2 = 9$, $s_X^2 = 16$, and $r_{xy} = 0.75$, what are the least squares estimates of b_0 and b_1? What is the R^2 from the least squares regression?

Solution.

$$R^2 = r_{xy}^2 = 9/16 = 0.5625, \quad b_1 = r_{xy}\frac{s_y}{s_x} = 0.75\frac{\sqrt{9}}{\sqrt{16}} = 0.5625, \quad b_0 = \bar{Y} - b_1\bar{X} = -6-(9/16)(4) = -8.25.$$

(c) Which of the following **always** results in a wider predictive interval for Y_f at a new location X_f? Circle all that apply.

(a)	a larger sample size (n)	(b)	a larger value of $\hat{Y}_f$
(c)	a larger degree of confidence (smaller α)	(d)	an X_f with lower leverage
(e)	a smaller estimated residual variance (s^2)	(f)	none of these

Solution. *Only choice (c) is correct. Refer to the formula at the end of the Lecture 3 slides. The smaller* α *leads to a larger quantile (either from the t distribution or the Normal), so that* $\hat{Y}_f \pm t_{n-p,\alpha} s_{pred}$ *is wider. In general, but not always, a larger sample size, an* X_f *with lower leverage, and a smaller* s^2 *will lead to narrower intervals. The value of* $\hat{Y}_f$ *itself has no effect.*

2 Understanding regression output #1

From the below `summary` of the regression of women's labor force participation (WLFP) in nineteen cities in 1972 (`wlfp72`) on WLFP in 1968 (`wlfp68`), answer the questions below.

```
Call:
lm(formula = wlfp72 ~ wlfp68)

Residuals:
     Min       1Q   Median       3Q      Max
-0.13086 -0.02797  0.01493  0.03678  0.06837

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.17435    0.09611   1.814  0.08736 .
wlfp68       0.60513    0.18088   3.345  0.00383 **
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05433 on 17 degrees of freedom
Multiple R-squared: 0.397,Adjusted R-squared: 0.3615
F-statistic: 11.19 on 1 and 17 DF,  p-value: 0.003835
```

(a) What is the t-statistic for a hypothesis test of whether or not the slope is equal to one? Write out the null and alternative hypotheses, and explain what the test means in terms of WLFP. What do you conclude at significance level $\alpha = 0.05$? Discuss if you think this conclusion is reasonable given the degrees of freedom.

Solution. *The null and alternative hypotheses are* $H_0 : \beta_1 = 1$ *vs.* $H_1 : \beta_1 \neq 1$*. We are trying to test if there is an increase or decrease in WLFP over the 4 years. Under the null,* $\beta_1 = 1$*, and the two years have the same WLFP. If we reject the null, we conclude that WLFP is lower, because* $\beta_1 \neq 1$ *and* $b_1 = 0.60513 < 1$*. The test statistic is*

$$\left| \frac{b_1 - \beta_1^0}{s_{b_1}} \right| = \left| \frac{0.60513 - 1}{0.18088} \right| = 2.183.$$

At level $\alpha = 0.05$*, we reject the null hypothesis, since this value is bigger than 2. This conclusion is reasonable if with 17 degrees of freedome we feel the Normal approximation is a good one. However, here we only have 17 degrees of freedom, and since the t distribution has fatter tails, we should be careful (*$t_{17,0.025} = 2.11$*, so it's still significant, but very close).*

(b) Suppose that the correlation between log(`wlfp72`) and log(`wlfp68`) was 0.67. Based on this information, would you describe the corresponding log-log model as a better fit? Why?

Solution. *You can not compare* R^2 *across different nonlinear transformations, so you can not conclude anything based on this information.*

3 True or False

For each question, circle either T (true) or F (false). Answering "true" implies that the given statement is *always* true. Statements are made in the context of this class, and the usual SLR/MLR assumptions.

(a) F Forecast uncertainty for Y_f does not depend on the input X_f.

(b) F A confidence interval for β_1 is centered at β_1.

(c) T Our linear regression model implies an error variance that is the same for all values of the explanatory variable.

(d) T Least squares residuals are not correlated with the fitted values.

(e) T All else being equal, a prediction interval is wider if the standard error for b_0 is larger.

(f) T Uncertainty about the regression coefficients depends upon the variance of the residuals.

(g) T The R^2 for a regression of Y onto X is the same as R^2 for the regression of X onto Y.

(h) T It is possible to reject a null hypothesis when the null hypothesis is true.

(i) F Our linear regression model implies that the marginal distribution for Y is normal.

(j) T Assuming our multiple linear regression model, each least squares coefficient b_j has mean β_j.

(k) F In simple linear regression, the slope of the regression line is equal to the correlation between X and Y.

(l) T Least squares estimates of the coefficients $\{b_0, b_1, ...,\}$ are chosen to maximize R^2.

4 Analyzing Plots

(a) There is nonconstant variance. The variance is nearly zero at $X = 1/2$, and grows in either direction. Dividing by $|X_i - 1/2|$ or $(X_i - 1/2)^2$ would help.

(b) The errors ε are correlated with X. Our SLR model assumes the two are independent. First, notice that the plotted line has an intercept of 1 and a slope of -2, matching the equation given for Y. But the errors around that line are different for different values of X, violating independence. In particular, the errors are mostly negative for small X and mostly positive for large X. The variance of the errors is constant, it is the value itself that is correlated with X. You can see this from the spread away from the line not changing with X, just which side of the line it is on.

5 Multiple Linear Regression 1: Electricity Demand

For an energy company in Alabama, we have the daily total electricity demand (measured in `MegaWatts`) and daily temperature (`temp` in degrees Fahrenheit above 32) for 364 days, with the day of the week (Sunday, Monday, ...) stored in `weekday`. Our goal is to predict electricity demand, so that the energy company can operate efficiently.

(a) Consider a regression of `MegaWatts` on the categorical `weekday`. Below are the output results from two different versions of this regression. In the first, Sunday is the baseline category, while in the second, Monday is the baseline. Answer the questions below the output.

Regression 1

```
Call:
lm(formula = MegaWatts ~ weekday)

Coefficients:
             Estimate Std. Err. t value Pr(>|t|)
(Intercept)      3162        62      51  <2e-16 ***
weekday2_Mon      288        87       3   0.001 **
weekday3_Tue      375        87       4   2e-05 ***
weekday4_Wed      345        87       4   9e-05 ***
weekday5_Thu      263        87       3   0.003 **
weekday6_Fri      278        87       3   0.002 **
weekday7_Sat      174        87       2   0.046 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 400 on 357 degrees of freedom
Multiple R-squared: 0.07,  Adjusted R-squared:  0.05
F-statistic:  4 on 6 and 357 DF,  p-value: 4e-04
```

Regression 2

```
Call:
lm(formula = MegaWatts ~ weekday)

Coefficients:
             Estimate Std. Err. t value Pr(>|t|)
(Intercept)      3451        62    56.1  <2e-16 ***
weekday2_Tue       87        87     1.0   0.318
weekday3_Wed       56        87     0.6   0.519
weekday4_Thu      -26        87    -0.3   0.767
weekday5_Fri      -10        87    -0.1   0.908
weekday6_Sat     -114        87    -1.3   0.189
weekday7_Sun     -288        87    -3.3   0.001 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 400 on 357 degrees of freedom
Multiple R-squared: 0.07,  Adjusted R-squared:  0.05
F-statistic:  4 on 6 and 357 DF,  p-value: 4e-04
```

(i) Which day of the week on average has the highest electricity demand? The lowest? Justify your answer numerically.

Solution. *The regression on the left has positive coefficients for all the other days, so Sunday must be lowest. The largest coefficient is for Tuesday, so it is the highest. The same conclusion follows from the right: Sunday has the most negative intercept, Tuesday the largest positive one.*

(ii) Discussing **both** of the regression outputs, what do you learn from the t tests and their associated p-values? Be specific and justify your answer numerically.

Solution. *The regression on the left has all dummies significant at the 5% level, so all the days are statistically different from Sunday. Regression 2 shows that Sunday is the only day statistically different from Monday. We can conclude some other things, in a more subtle way: notice the standard errors don't change depending on the baseline day, and that a coefficient of -114 is not significant (Saturday in regression 2), which implies that any two days that are less than 114 apart will not be significantly different from each other. For example, Saturday is not different than Friday, nor is Tuesday statistically significantly different from Friday. In the same way, Tuesday and Saturday are 114+87 = 201 apart, and regression 1 shows that a difference of 174 is significant, and so 201 must be as well. Saturday and Wednesday are 170 apart, which may or may not be statistically significant, we don't know.*

(iii) Using **both** of the about regression outputs, what do you learn from the F test? Conceptually, why do the two regressions have the same F test?

Solution. *From the F-test we learn that (statistically) at least one coefficient is different from zero, and hence at least one day is different from another (the baseline day). The two regressions are identical up to coding the day variable differently, and hence they have the same difference or similarity. That is, these coefficients explain a fixed amount of the variation, no matter how you permute them.*

(b) Using the output from the model below, which includes `temp` and `weekday`, what is the predicted total electricity demand (measured in MegaWatts) for a Wednesday where the temperature is 52 degrees Fahrenheit? If the standard error of the fitted value is $s_{\text{fit}} = 300$ MegaWatts, what is a 95% predictive interval for your answer?

```
Call:
lm(formula = MegaWatts ~ weekday * temp)

Coefficients:
```

```
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)          2632        243      11   <2e-16 ***
weekday2_Tue          495        345       1    0.153
weekday3_Wed         -476        381      -1    0.212
weekday4_Thu        -1352        416      -3    0.001 **
weekday5_Fri        -1012        401      -2    0.012 *
weekday6_Sat        -1001        398      -2    0.012 *
weekday7_Sun        -1330        386      -3    6e-04 ***
temp                   19          5       3    7e-04 ***
weekday2_Tue:temp      -9          8      -1    0.225
weekday3_Wed:temp      12          8       1    0.173
weekday4_Thu:temp      29          9       3    0.002 **
weekday5_Fri:temp      22          9       2    0.014 *
weekday6_Sat:temp      19          9       2    0.030 *
weekday7_Sun:temp      23          9       3    0.007 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 400 on 350 degrees of freedom
Multiple R-squared:  0.4,       Adjusted R-squared:  0.4
F-statistic: 2e+01 on 13 and 350 DF,  p-value: <2e-16
```

Solution. *The tricky part here is that* `temp` *is measured as degrees above 32, and hence if it is 52 degrees, our prediction is based on* `temp` *= 20. Using this and the fact that it is Wednesday, we find that*

$$\hat{Y} = 2632 - 476 + (19 + 12) \times 20 = 2776.$$

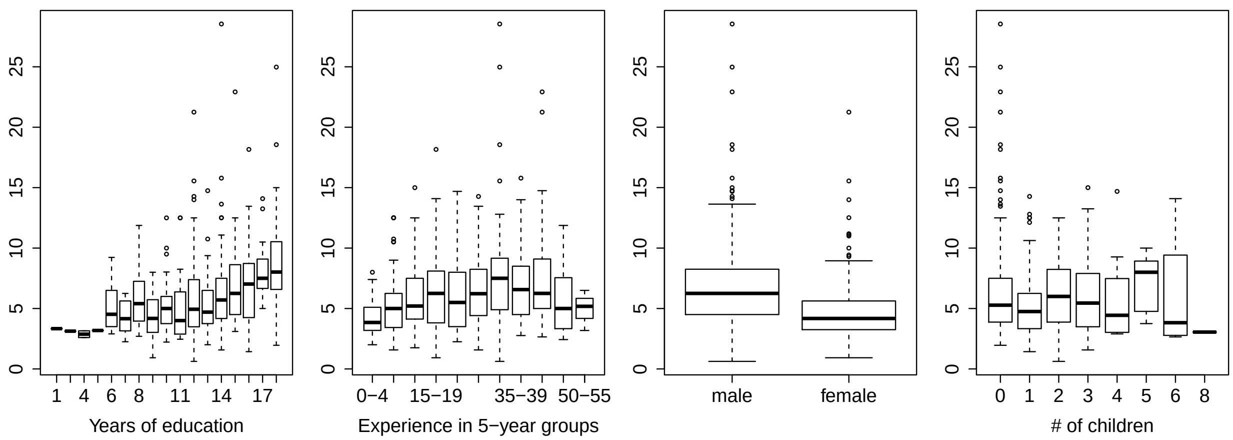
The key is to not forget the additional slope due to the interaction term. Recalling that $s_{pred} = \sqrt{s^2 + s_{fit}^2} = \sqrt{400^2 + 300^2} = 500$ *(the former from the output, the latter is given) a 95% predictive interval is* $\hat{Y} \pm 2 \times 500$.

6 Multiple Linear Regression 2: Predicting Wages

This problem examines predicting wages based on observed characteristics. The data consists of 550 employed individuals in 1978 and has the following variables:

`wage` = Hourly wage in 1978
`educ` = Years of education completed
`exper` = years of labor market experience
`female` = 1 if female, 0 if male
`kids` = number of dependent children.

(a) Comment on the relationship between `wage` and the other four variables using the boxplots below. Note any patterns as well as concerns.



Solution. *(1) Wage increases with education, but there is clearly much more variability with higher levels of education; note that for less than 6 years of education there is virtually zero variance. (2) Wage increases with experience, but then appears to decrease at very high levels, indicating possible nonlinearity. (3) Females earn lower wages, on average. (4) Children appears to have no effect, on average. I would be concerned that very few people have more than 4 children, and so the estimates may have outlier/leverage type problems. There are a lot of plausible stories that can go along with these.*

Consider the following `summary`.

```
Call:
lm(formula = log(wage) ~ educ + exper + female + kids)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.39318 -0.25112  0.02287  0.24630  1.32295 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.590007   0.099703   5.918 5.78e-09 ***
educ         0.077766   0.006601  11.781  < 2e-16 ***
exper        0.013355   0.001378   9.694  < 2e-16 ***
female      -0.337317   0.035669  -9.457  < 2e-16 ***
kids        -0.007021   0.013411  -0.523    0.601    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4007 on 545 degrees of freedom
Multiple R-squared:  0.3367,    Adjusted R-squared:  0.3318 
F-statistic: 69.15 on 4 and 545 DF,  p-value: < 2.2e-16
```

(b) Give a precise, numerical interpretation of what the above output says about the association between `educ` and `wage`.

Solution. *First, the p-value is very small, and hence education does significantly predict wage, holding fixed experience, sex, and number of children. To fully answer the question, we have to undo the log transformation. If*

$$\log(\texttt{wage}) = \beta_0 + \beta_1\texttt{educ} + \beta_2\texttt{exper} + \beta_3\texttt{female} + \beta_4\texttt{kids}$$

then

$$\texttt{wage} = e^{\beta_0} \times e^{\beta_1\texttt{educ}} \times e^{\beta_2\texttt{exper}} \times e^{\beta_3\texttt{female}} \times e^{\beta_4\texttt{kids}}.$$

Therefore

$$\frac{\partial \texttt{wage}}{\partial \texttt{educ}} = \beta_1 e^{\beta_0} \times e^{\beta_1 \texttt{educ}} \times e^{\beta_2 \texttt{exper}} \times e^{\beta_3 \texttt{female}} \times e^{\beta_4 \texttt{kids}} = \beta_1 \texttt{wage} \Rightarrow \beta_1 = \frac{\partial \texttt{wage}}{\partial \texttt{educ}} \frac{1}{\texttt{wage}} = \%\ \textit{change in}\ \texttt{wage}.$$

So we conclude that an extra year of education is associated with a 7.78% increase in hourly wage. The last important point is that this is holding everything else equal, i.e. conditional on (controlling for) everything else in the model. You could say something like, "after taking into account the effects of experience, sex, and number of children, an 8% wage increase is associated with each additional year of education".

(c) Using the summary above, give a 95% confidence interval for the coefficient of `kids`. Interpret your answer, referring to part **(a)**.

Solution. *The confidence interval is* $[-0.007021 \pm 2 \times 0.013411] = [-0.0338, 0.019801]$. *The confidence interval includes zero, so we can not rule out that the coefficient is zero, which is exactly as it appeared in part (a): there appears to be no linear trend relating wage and number of children.*

(d) Define `exper.squared` = (`exper`)^2. Using the summary below, should `exper.squared` be included in the model? Why or why not? Interpret your answer, referring to part **(a)**.

```
Call:
lm(formula = log(wage) ~ educ + exper + exper.squared + female +
    kids)

Residuals:
     Min       1Q   Median       3Q      Max
-2.42121 -0.23810  0.01701  0.24124  1.37470

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.5615258  0.0988398   5.681 2.18e-08 ***
educ           0.0719948  0.0067054  10.737  < 2e-16 ***
exper          0.0318576  0.0051465   6.190 1.18e-09 ***
exper.squared -0.0004211  0.0001130  -3.728 0.000213 ***
female        -0.3419768  0.0352764  -9.694  < 2e-16 ***
kids          -0.0285579  0.0144596  -1.975 0.048772 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.396 on 544 degrees of freedom
Multiple R-squared:  0.3532,    Adjusted R-squared:  0.3472
F-statistic: 59.41 on 5 and 544 DF,  p-value: < 2.2e-16
```

Solution. *Yes,* `exper.squared` *should be included. The p-value is* $0.000213 \ll 0.05$, *and so this term aids the model fit. From part (a), we can see the nonlinearity in the histrograms: for high values of experience, wage decreases again, there is a slight parabolic shape. The negative slope estimate for* `exper.squared` *makes sense in this light. The* R^2 *has barely increased, but that doesn't matter here:* R^2 *always goes up when you include more variables.*

7 Regression: Baseball Data

For each Major League Baseball team we have the number of wins (`Wins`) and the total player salary in millions of dollars (`Salary`) for 2006. (You don't need to know anything about baseball for this question.) The total league payroll was $2,326.707 million. For each team i, define

$$\texttt{SalaryShare}_i = \frac{\texttt{Salary}_i}{\sum_{j=1}^{n} \texttt{Salary}_j} = \frac{\texttt{Salary}_i}{2,326.707}.$$

Now consider the following `summary`.

```
Call:
lm(formula = Wins ~ SalaryShare)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.7907  -4.5503   0.3654   4.5352  17.4042 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   67.982      4.178  16.271  8.4e-16 ***
SalaryShare  389.540    116.013   3.358  0.00228 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 8.665 on 28 degrees of freedom
Multiple R-squared:  0.2871,     Adjusted R-squared:  0.2616 
F-statistic: 11.27 on 1 and 28 DF,  p-value: 0.002277
```

(a) But suppose that instead of regressing `Wins` on `SalaryShare` we used `Salary` itself as the input. Use the summary above to compute the estimates of the intercept b_0, the slope b_1, and the R^2 value for this hypothetical regression.

Solution. *The model being run for the output above is that*

$$\texttt{Wins}_i = b_0 + b_1\texttt{SalaryShare}_i + e_i = b_0 + b_1\frac{\texttt{Salary}_i}{2,326.707} + e_i = b_0 + \frac{b_1}{2,326.707}\texttt{Salary}_i + e_i.$$

Therefore, if we regress `Wins` *on just* `Salary`*, we will get the same* b_0*. The new slope estimate will be* $= b_1/2,326.707 = 0.167$*. The residuals will also be the same, and hence the* R^2 *will be the same. Intuitively, we are just changing the scale that we measure salary. It's exactly the same as changing prices from dollars to thousands of dollars: here instead of dividing by \$1,000, we are dividing by \$2,326.707 million, but it's the same idea. Changing the scale doesn't change how two variables are associated in the data.*

(b) Do we have reason to believe in a linear relationship between `Wins` and `Salary`, in the hypothetical regression in part **(b)**? State a formal hypothesis test, the value of the test statistic, and the conclusion.

Solution. *The null and alternative hypotheses are* $H_0 : \beta_1 = 0$ *vs.* $H_1 : \beta_1 \neq 0$ *in the model* $\texttt{Wins} = \beta_0 + \beta_1\texttt{Salary} + \varepsilon$*. The value of the test statistic is 3.358, from above, which is greater than 2, and thus significant at the 0.05 level. Why is that the test statistic? That number, from the output is for testing whether the slope on* `SalaryShare` *is equal to zero. Referring to the previous answer, if* $\beta_{\texttt{SalaryShare}}$ *is zero, then certainly* $\beta_{\texttt{SalaryShare}}/2,326.707 = \beta_{\texttt{Salary}}$ *is zero as well, and vice versa. So testing if one is zero is identical to testing the if the other is zero. A more mechanical way to see this is to recompute the test statistic. To compute the standard error:*

$$s^2_{b_{\texttt{Salary}}} = \mathbb{V}\left[\frac{b_{\texttt{SalaryShare}}}{2,326.707}\right] = \frac{\mathbb{V}[b_{\texttt{SalaryShare}}]}{2,326.707^2} = \frac{s^2_{b_{\texttt{Salary}}}}{2,326.707^2}.$$

Therefore

$$\left|\frac{b_{\texttt{Salary}} - 0}{s_{b_{\texttt{Salary}}}}\right| = \left|\frac{b_{\texttt{SalaryShare}}/2,326.707}{s_{b_{\texttt{Salary}}}/2,326.707}\right| = \left|\frac{b_{\texttt{SalaryShare}} - 0}{s_{b_{\texttt{Salary}}}}\right|,$$

which is exactly what is reported in the output.

(c) In 2006 the Chicago White Sox payroll was \$102.75 million and won 90 games. What is the predicted number in wins if they added \$10 million to their payroll? If the standard error of the fitted value is $s_{\text{fit}} = 2$, what is a 95% predictive interval for your answer?

Solution. *First, note that the number of games they actually won is totally irrelevant to predicting. We are going to read the prediction off the line, this is only the same as the observed outcome if the* *residual is zero. As usual, only one or two points in the scatter plot have a residual of zero. Our aim is to predict at a new value, the old values don't matter. The question isn't about the change in number of wins, or the increase. In fact, we'll see that the predicted value is below 90, which means that the regression line lies below this point.*

The key is that we need to take into account the change in total salary. The easiest way to get this answer is from the `SalaryShare` *regression output. The key part to realize is that the total salary will increase as well. The fitted value is*

$$\hat{Y} = 67.982 + 389.540\frac{102.75 + 10}{2,326.707 + 10} = 86.778.$$

The predictive interval is

$$\hat{Y} \pm 2 \times \sqrt{2^2 + 8.665^2} = 86.778 \pm 2 \times 8.89 = [68.99, 104.56].$$

It is tempting, but wrong, to use the results of the hypothetical regression of `Wins` *on* `Salary`*:* $\hat{Y} = b_0 + b_{\texttt{Salary}} 112.75 = 67.982 + 0.1674 \times 112.75 = 86.86$. *Whoops! This answer is very close, but it is slightly wrong because it fails to take into account the change in total salary, and hence the change in scaling. The computation of* $b_{\texttt{Salary}} = 0.1674$ *assumed a scaling of 2,326.707. In fact, if you use the* `SalaryShare` *regression but plug in* $X_f = \frac{102.75+10}{2,326.707}$ *you will get this same, incorrect prediction. To get the right answer, you have to "deflate" the new* `Salary` *by the change in total salary. Remember that* $b_{\texttt{Salary}} = b_{\texttt{SalaryShare}}/2,326.707$

$$\hat{Y} = b_0 + b_{\texttt{SalaryShare}}\frac{102.75 + 10}{2,326.707 + 10} = b_0 + \underbrace{\frac{b_{\texttt{SalaryShare}}}{2,326.707}}_{b_{\texttt{Salary}}}\underbrace{\left[\frac{2,326.707}{2,326.707 + 10}(102.75 + 10)\right]}_{\textit{New Salary}} = 86.778.$$

The term in square brackets is the "New Salary" with the new total, and equals 112.267, slightly less than 112.75. The idea is that having a payroll of 112.267 when the total salary is 2,326.707 is the same as having a payroll of 112.75 when the total is 2,326.707 + 10. The interval can be computed similarly.